NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

A TEXT ANALYSIS OF THE MARINE CORPS FITNESS REPORT

by

Philipp E. D. Rigaut

June 2017

Thesis Advisor:	Robert Koyak
Second Reader:	Chad Seagren

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** June 2017	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
A TEXT ANALYSIS OF THE MARINE CORPS FITNESS REPORT

5. FUNDING NUMBERS

6. AUTHOR(S) Philipp E. D. Rigaut

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)
USMC Manpower and Reserve Affairs
3280 Russell Rd
Quantico, VA 22134-5103

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB number NPS.2011.0007-rR-EPS.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

The Marine Corps loses about half of its nearly two thousand officers at the end of their initial contracts for various reasons. In an effort to control talent retention, the Marine Corps is examining if the appropriate evaluation structure is in place to identify the top performers. This study is an analysis of textual information contained in fitness reports to determine the extent to which it informs promotion boards of the quality of a Marine officer. We examine 71,212 observed fitness reports from the 1996, 1997, 2006, and 2007 officer cohorts, which we observe from 2007 to 2016. We use text statistics, readability indicators, natural language processing, and a variety of statistical machine learning algorithms to predict the top and bottom performers. We find that fitness reports for the best-performing officers are well written, use simple words in longer sentences, and comment on future command opportunities. Remarks on performance, potential, billet assignment, and education do not contribute predictive power. The fitness report contributors often disagree and informative power is lost when the assigned marks do not conform to issued guidance. In isolation, the comment sections are inconclusive for predicting an officer's performance tier. We attain a correct classification rate of 67% when using an optimized ensemble of prediction models. We recommend that the Marine Corps provide word-picture guidance to distinguish talented Marines and promote conformity in issuing quantitative assessments of performance.

14. SUBJECT TERMS
natural language processing, fitness reports, computational linguistics, manpower

15. NUMBER OF PAGES 197

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

A TEXT ANALYSIS OF THE MARINE CORPS FITNESS REPORT

Philipp E. D. Rigaut
Major, United States Marine Corps
B.S., United States Naval Academy, 2005

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2017

Approved by: Robert Koyak
Thesis Advisor

Chad Seagren
Second Reader

Patricia A. Jacobs
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Marine Corps loses about half of its nearly two thousand officers at the end of their initial contracts for various reasons. In an effort to control talent retention, the Marine Corps is examining if the appropriate evaluation structure is in place to identify the top performers. This study is an analysis of textual information contained in fitness reports to determine the extent to which it informs promotion boards of the quality of a Marine officer. We examine 71,212 observed fitness reports from the 1996, 1997, 2006, and 2007 officer cohorts, which we observe from 2007 to 2016. We use text statistics, readability indicators, natural language processing, and a variety of statistical machine learning algorithms to predict the top and bottom performers. We find that fitness reports for the best-performing officers are well written, use simple words in longer sentences, and comment on future command opportunities. Remarks on performance, potential, billet assignment, and education do not contribute predictive power. The fitness report contributors often disagree and informative power is lost when the assigned marks do not conform to issued guidance. In isolation, the comment sections are inconclusive for predicting an officer's performance tier. We attain a correct classification rate of 67% when using an optimized ensemble of prediction models. We recommend that the Marine Corps provide word-picture guidance to distinguish talented Marines and promote conformity in issuing quantitative assessments of performance.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

1stLt	first lieutenant
2ndLt	second lieutenant
A-PES	Advanced-Performance Evaluation System
ARI	Automated Readability Index
BMOS	billet military occupational specialty
Capt	captain
CLI	Coleman-Liau index
CLS	career level school
CMC	Commandant of the Marine Corps
CNA	Center for Naval Analyses
CDF	cumulative distribution function
DTM	document-term matrix
ECDF	empirical cumulative distribution function
FITREP	fitness report
FRA	fitness report averages
Glm	generalized linear model
HPC	higher performance computing
Elastic net	Lasso and Elastic-Net Regularized Generalized Linear Models
IRAM	Individual Records Administration Manual
ILS	intermediate level school
IS	Islamic State
LtCol	lieutenant colonel
M&RA	Manpower and Reserve Affairs
Maj	Major
Maxent	maximum entropy
MCO	Marine Corps Order
MOS	military occupational specialty
MRO	Marine reported on
MARADMIN	Marine administration
MARCORPROMMAN	Marine Corps Promotions Manual

MARCORSEPMAN	Marine Corps Separations Manual
MBS	master brief sheet
MMRP	Manpower Management Records and Performance
NAVMC	Navy Marine Corps
NLP	natural language processing
NPS	Naval Postgraduate School
Nnet	neural network
OAD	Operations Analysis Division
OCR	optical character recognition
OMPF	official military personnel file
ORM	operational risk management
PES	performance evaluation system
PERB	performance evaluation review board
PME	professional military education
RS	reporting senior
RO	review officer
RF	random forest
RV	relative value
SECNAVINST	Secretary of the Navy Instructions
SMOG	Simple measure of gobbledygook
SNO	said named officer
Svm	support vector machine
TF-IDF	term frequency-inverse document frequency
TDM	term-document matrix
TBS	The Basic School
TFDW	Total Force Data Warehouse
TLS	top level school
USMC	United States Marine Corps

EXECUTIVE SUMMARY

The Marine Corps loses about half of its nearly two thousand officers at the end of their initial contract. In an effort to control talent retention, the Marine Corps is examining whether the appropriate evaluation structure is in place to identify the top performers. This study is an analysis of textual information contained in fitness reports to determine the extent to which it informs readers of the quality of a Marine officer.

We examine 71,212 observed fitness reports on 4,761 officers commissioning in 1996, 1997, 2006, and 2007. We use text statistics, readability indicators, natural language processing, and a variety of statistical machine learning algorithms to predict the top, middle, and bottom performers from the text in those reports. In our thesis we use words, or statistics describing words, as predictors of a Marine officer's performance tier determined from the reporting senior's (RS) relative value and reviewing officer's (RO) comparative assessment.

We begin our analysis by inspecting the quality of our response variable: the Marine officer's performance tier. We find that the relative value distribution is susceptible to outliers and tends to concentrate in a narrow range. The density and measures of central tendency of the relative values exhibit an increasing trend with rank. The reviewing officer's comparative assessment tends to concentrate in a narrower range than the prescribed "Christmas tree" distribution, which makes it more difficult to distinguish performance among Marine officers. We find that the comparative assessments also have an increasing trend with rank. By assessing concurrence between the RS and RO evaluations, we find that although ROs rarely indicate formal non-concurrence with the RSs' evaluations, their assessments disagree 49% of the time. We conclude that relative value and comparative assessment tier groups are not precise measures of performance.

We search for informational value in the text fields of the fitness report body. These fields provide amplifying information on the performance and potential of the Marine officer under review. The reporting senior comments fall into three categories:

mandatory, directed, and additional comments. The reviewing officer comments on the administrative correctness of the fitness report and compares the Marine to others in the same grade.

We derive a set of metrics of writing quality: spelling errors, word counts, character counts, and five different assessments of readability. Fitness reports are commensurate with literature on product reviews and professor evaluations: well-written, simple words in longer sentences with few spelling mistakes are associated with positive sentiment. Although each of the metrics is a statistically significant predictor of a Marine’s performance, only word counts and readability indices that focus on character density are informative. We compare our models to a naïve selection, defined as the equal probability of selecting any tier. We develop two models that when combined correctly classify a Marine officer into his or her performance tier group 55% of the time or 22% better than naïve selection.

In the next phase of our analysis, we examine correlations of performance characteristics with keywords, directed comments, and comparative superlatives. Using supervised and unsupervised correlation models with syntagmatic word association, we find that language proximate to “promotion,” “potential,” “education,” and “assignment” do not exhibit predictive power due to the common occurrence of these terms across all performance tiers. Comments that note future command opportunities tend to indicate top-tier Marine officers, and the appearance of “peer” in the comments is associated with lower-tier performance. We find that directed comments are often absent when not prompted, but usually are provided when there is a reminder prompt. Interestingly, stating a Marine was the “top performer,” “number 1,” or similar constructs does not add predictive value. We conclude that when reading the comment fields, a reader gains marginal informational value in the textual body.

In the final stage of our analysis, we construct an optimal configuration of machine learning models to predict the performance tier. We recognize 360 different collections of word configurations specific to rank, tier, and the writer of the comment (reporting senior or reviewing officer), with word configurations varying in length from one to six. We use seven different supervised machine learning algorithms to find the

optimal bag-of-words configuration to stack into an ensemble of models. We find that frequency weighting of single words in penalty-enhance generalized linear models, support vector machines, classification and regression trees, maximum entropy models, and random forests provides the most predictive power. Together, they correctly classify 56% of Marines into performance tiers. When combined with the two writing-quality models, we improve our classification rate to 67%.

Throughout the study, we derive informational value from descriptive statistics, readability indices, keyword correlation, and supervised predictive modeling. Individually, these techniques provide slightly better correct classification rates than unskilled assignment; however, by creating an ensemble of multiple models in specific configurations, we double the correct classification rate. These correct classification rates are low compared to similar sentiment analysis in product reviews and professor evaluations. Our results can inform the Marine Corps on the use of language in fitness reports with the aim of adopting standardized language to ensure that quality of a Marine is consistently represented in fitness reports comments. We recommend that the Marine Corps enhance the word-picture guidance to separate talented Marines and promote conformity in issuing quantitative assessments of performance.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

My journey through completion of Naval Postgraduate School was long and rewarding, but I could not have done it by myself. Throughout my life, many people have influenced my perception and have greatly contributed to my professional and personal success. I first and foremost want to thank God for providing for and watching over my family and me. I especially want to thank my wonderful family, Mandi, Drew, Laila, Emory, Muff, and Chloe, who have always been by my side. A fair amount of people have directly or indirectly influenced me by dedicating a considerable amount of time, effort, or ideas in my development. I would like to thank Robert Davies, John Meixner, Chris Medlin, Doug Cromwell, Jay Lyn, Adam Tahir, Chris Frey, and Marcus Mainz.

My time at NPS was extremely pleasant, and I am amazed at the professionalism and sheer brain power the faculty possess. I'd like to thank those that greatly enhanced my séjour here at NPS: Professors Beverly, Carlyle, Dell, Lucas, Buttery, Whitaker, Sampson, Lin, and Atkinson. I am also deeply thankful to the 2017 Marine cohort: Scott, Jimmy, Dare, Elle, Kat, Sam, Booby, Micah, Ezra, Zach, and Taylor. What an incredible group of talented, selfless officers. I couldn't have done it without you. Finally, I would like to thank those who helped me complete this thesis: Professors Whitaker, Koyak, Seagren, and Doreen Marucci, and the great crew of OR graduates at M&RA.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

When General Robert Neller assumed his duties as the Commandant of the Marine Corps on September 24, 2015, he declared his highest priority as taking care of the people who comprise the Marine Corps. In an effort to sustain the Marine Corps's combat effectiveness, Neller mentioned in the necessity of "maintaining a force of the highest quality, which is smart, resilient, fit, disciplined, and able to overcome adversity. Recruiting and retaining quality men and women of character in today's Corps is our friendly center of gravity and our highest priority" (Neller, 2016).

The Marine Corps's largest expense is manpower and its associated functions (Department of Defense, 2015). In a fiscally constrained environment, the screening, retention, and promotion of Marines play essential roles in meeting the aforementioned objectives. In order to measure Marines' performance and potential for future service, the fitness report serves as the most important information component in manpower management. The Commandant views the fitness report as the "primary tool available for the selection of personnel for promotion, retention, augmentation, resident schooling, command, and duty assignments" (Commandant of the Marine Corps, 2015). The Marine Corps Separations Manual (MARCORPROMMAN) provides the following guidance: "Officers are selected for promotion for their potential to carry out the duties and responsibilities of the next higher grade based upon past performance as indicated in their official military personnel file" (Commandant of the Marine Corps, 2006).

The current fitness report system has been in place since 1999. To determine whether the system is effective, the Director of the Manpower Management Division at Manpower and Reserved Affairs commissioned the Center for Naval Analyses (CNA) to conduct a survey in order to determine whether the system was fair and whether the various boards selected the "best and most qualified" officers (Clemens et al., 2012). The study offers a cross-sectional analysis of reporting senior and reviewing officer reporting tendencies across a 13-year span and concludes that the system appears to be effective. The CNA study did, however, mention some certain critical areas of concern regarding community and personal attribute bias.

A. PROBLEM STATEMENT

An increasing number of studies over the last 20 years have examined the Marine Corps fitness reports and the performance evaluation system. These include a collection of Naval Postgraduate School master's thesis and the study mentioned above (Phillips & Clemens, 2011; Clemens, Malone, Phillips, & Lee, 2012; Ergun, 2003; Jobst & Palmer, 2005; Reynolds, 2011; Gonzales, 2012). These studies, however, have not considered comments under Sections I and K due to technological limitations and availability of information. Recent advances of both computational power and the development of more sophisticated text mining techniques have allowed more detailed analyses to be performed on a larger sample. The comments are largely unstructured data, which makes it difficult to process and analyze them.

The focus of analysis for this study is to show whether there is a quantifiable correlation between the trait-scaled response evaluation and the results of supervised learning models on the reporting senior and reviewing officer comments. If there is a significant correlation, it could have a significant impact on the way that reporting senior (RS) and reviewing officer (RO) comments are designed.

B. OBJECTIVES

The purpose of this paper is to analyze the informational value of the text fields in a fitness report. In fulfilling this purpose, we address the following research questions:

1. What is the relationship between the relative value and attributes of the section I comments recorded by the reporting senior on a fitness report?

2. What is the relationship between the comparative assessment and attributes of the section K comments recorded by the reviewing officer on a fitness report?

To answer these questions, our analysis proceeds as follows:

1. Analyze the relative value and comparative assessment distributions.

2. Define readability metrics of the comment fields through.

    - Comment lengths

    - Word counts

- Spelling errors
- And five readability indices

3. Examine contents of the comment field.
4. Examine the relationship between RS markings and comments.
5. Examine the relationship between RO markings and comments.

C. SCOPE AND LIMITATIONS

1. Nature of the Study

To our knowledge, our thesis is the first analysis that considers the textual information in fitness reports. Previous research focuses on fitness report markings to express relationships between attributes of the MRO, RS, and RO; and categorical variables that include demographics and various indicators of performance (Clemens et al., 2012). Jobst and Palmer (2005) examine fitness reports to determine whether entry-level performance and commissioning source are useful indicators of performance. Gonzales (2012) investigates promotability of eligible officers to lieutenant colonel based on the relative rankings in fitness reports. Reynolds (2011) uses only summary relative values from fitness reports in conjunction with categorical variables taken from other sources, such as military education, combat experience, education above bachelor's degree. None of these prior studies address sections I and K comments in relations to the assigned markings. There are a wide range of issues that arise when discussing evaluations but we focus on quantifying the relationship between the markings and the comments.

2. Methodology

We analyze the textual value of the fitness report by considering the whole fitness report. Focusing on the performance tier classification system, we investigate the quality of our response variable. By investigating textual descriptive statistics, we assess the readability of fitness reports and its value in predicting tier groups. Using a modified adaption of a text mining workflow (Jordan, 2011, p. 49), we process the body of fitness reports using Freinere and Hornik's tools (2015) and a family of supervised machine

learning models implemented by Jurka, Collingwood, Boydstun, Grossman, and van Atteveldt (2012) to create *k* best predictors of *n*-word combinations for comparison with the scaled responses.

D. ORGANIZATION OF THE THESIS

The study is organized into five chapter with five appendices. Chapter II details the fitness report, outlines the education officers receive on writing them, previous studies in performance evaluations, and natural language processing literature review. Chapter III proposes our process and methodology for converting the data into useful model inputs. We detail our approach and some mathematical tools we use in our supervised machine learning models. Chapter IV presents our results separated into four parts: analysis on the quality of our response variable, descriptive statistics on the fitness report comments, analysis of the keywords, and predictive modeling. Chapter V concludes our study, summarizes our results, and provides recommendations.

II. BACKGROUND

A. THE USMC ADVANCED PERFORMANCE EVALUATION SYSTEM

The USMC advanced performance evaluation system (A-PES), was implemented in January 1999 and designed to reduce grade inflation (Clemens, Malone, Phillips, & Lee, 2012, p. 11; Phillips & Clemens, 2011, pp. 1,12). While the current fitness report system retains almost all the administrative features that “guide the preparation and submission of reports” (Ergun, 2003, p. 28), it adds duties, responsibilities, and accountability of the RS and RO (Commandant of the Marine Corps, 2015, pp. 2-1 to 2-5). The format includes a seven-point ordinal rating scale, a reduced emphasis for word picture comments, use of the relative value; and it eliminates relative comparisons among peers by the reporting senior (Ergun, 2003). The following paragraphs discuss each section of the current format.

The A-PES provides the framework to capture critical performance metrics under 12 sections lettered from A to L. Section A comprises all the administrative information that places the MRO in time and billet. Section A also contains comments on mandatory annual Marine Corps requirements that factor into promotion such as physical fitness and marksmanship scores. Section A also contains special remarks on applicable adverse, commendatory, or derogatory material. When any special fields are marked, the RS must make a corresponding directed comment in Section I.

Section B provides the reporting senior an “opportunity to describe the scope of duties which form the basis for evaluating” (Commandant of the Marine Corps, 2015, p. 4-17). Section C contains information on what the MRO accomplished during the reporting period (Commandant of the Marine Corps, 2015, p. 4-19). Under A-PES, the MRO has the opportunity to enhance the RS’s and RO’s perceptions of billet accomplishments by providing an “MRO worksheet.” This worksheet allows the MRO to provide administrative information and his or her perception of significant accomplishments in Section C.

Sections D through H highlight the 14 most important attributes to evaluate in the performance of a Marine. These qualities are subdivided into five sections: mission accomplishment, individual character, leadership, intellect and wisdom, and fulfillment of evaluation responsibilities (Commandant of the Marine Corps, 2015, p. 4-21). Collectively, “these attributes provide a clear picture of a Marine’s demonstrated capacities, abilities, and character” (Commandant of the Marine Corps, 2015, p. 4-21). The following attribute descriptions are taken verbatim from the fitness reports form and are used by the Reviewing Senior to qualify the MRO’s performance:

> **Performance**. Results achieved during the reporting period. How well those duties inherent to a Marine’s billet, plus all additional duties, formally and informally assigned, were carried out. Reflects a Marine’s aptitude, competence, and commitment to the unit’s success above personal reward. Indicators are time and resource management, task prioritization, and tenacity to achieve positive ends consistently.
>
> **Proficiency**. Demonstrates technical knowledge and practical skill in the execution of the Marine’s overall duties. Combines training, education and experience. Translates skills into actions which contribute to accomplishing tasks and missions. Imparts knowledge to others. Grade dependent.
>
> **Courage**. Moral or physical strength to overcome danger, fear, difficulty or anxiety. Personal acceptance of responsibility and accountability placing conscience over competing interests regardless of consequences. Conscious, overriding decision to risk bodily harm or death to accomplish the mission or save others. The will to persevere despite uncertainty.
>
> **Effectiveness Under Stress**. Thinking, functioning and leading effectively under conditions of physical and/or mental pressure. Maintaining composure appropriate for the situation, while displaying steady purpose of action, enabling one to inspire others while continuing to lead under adverse conditions. Physical and emotional strength, resilience and endurance are elements.
>
> **Initiative**. Action in the absence of specific direction. Seeing what needs to be done and acting without prompting. The instinct to begin a task and follow through energetically on one’s own accord. Being creative, proactive and decisive. Transforming opportunity into action.
>
> **Leading Subordinates**. The inseparable relationship between leader and led. The application of leadership principles to provide direction and motivate subordinates. Using authority, persuasion and personality to

influence subordinates to accomplish assigned tasks. Sustaining motivation and morale while maximizing subordinates' performance.

Developing Subordinates. Commitment to train, educate, and challenge all Marines regardless of race, religion, ethnic background, or gender. Mentorship. Cultivating professional and personal development of subordinates. Developing team players and esprit de corps. Ability to combine teaching and coaching. Creating an atmosphere tolerant of mistakes in the course of learning.

Setting the Example. The most visible facet of leadership: how well a Marine serves as a role model for all others. Personal action demonstrates the highest standards of conduct, ethical behavior, fitness, and appearance. Bearing, demeanor, and self-discipline are elements.

Ensuring Well-being of Subordinates. Genuine interest in the well-being of Marines. Efforts enhance subordinates' ability to concentrate/focus on unit mission accomplishment. Concern for family readiness is inherent. The importance placed on welfare of subordinates is based on the belief that Marines take care of their own.

Communications Skills. The efficient transmission and receipt of thoughts and ideas that enable and enhance leadership. Equal importance given to listening, speaking, writing, and critical reading skills. Interactive, allowing one to perceive problems and situations, provide concise guidance, and express complex ideas in a form easily understood by everyone. Allows subordinates to ask questions, raise issues, and concerns and venture opinions. Contributes to a leader's ability to motivate as well as counsel.

Professional Military Education (PME). Commitment to intellectual growth in ways beneficial to the Marine Corps. Increases the breadth and depth of warfighting and leadership aptitude. Resources include resident schools; professional qualifications and certification processes; nonresident and other extension courses; civilian educational institution coursework; a personal reading program that includes (but is not limited to) selections from the Commandant's Reading List; participating in discussion groups and military societies; and involvement in learning through new technologies.

Decision Making Ability. Viable and timely problem solution. Contributing elements are judgment and decisiveness. Decisions reflect the balance between an optimal solution and a satisfactory, workable solution that generates tempo. Decisions are made within the context of the commander's established intent and the goal of mission

accomplishment. Anticipation, mental agility, intuition, and success are inherent.

Judgment. The discretionary aspect of decision making. Draws on core values, knowledge, and personal experience to make wise choices. Comprehends the consequences of contemplated courses of action.

Evaluations. The extent to which this officer serving as a reporting official conducted, or required others to conduct, accurate, uninflated, and timely evaluations. (Commandant of the Marine Corps, 2015, Appendix B)

The seven markings of “A” to “H” correspond to three descriptions under each trait to guide the reporting senior into selecting the proper evaluation. An “A” is the lowest mark. It denotes unsatisfactory performance and renders the entire report adverse. On the other hand, “F” and “G” marks are the highest marks and express exceptional performance. Any of these three markings require justification by the RS in the space provided at the bottom of each section (Commandant of the Marine Corps, 2015 p. 4–22). Each letter corresponds to a numerical score ranging from 1 for “A” to 7 for “G.” When a specific trait is not observed or the RS cannot form an accurate assessment during the reporting period, the RS marks “H,” which reduces the denominator of the average by the number of un-observed traits. These numerical scores contribute to the raw score value of the fitness report by averaging the observed scores.

Section I, which is also known as the “word picture,” provides the RS an opportunity to expand on the performance and character of the evaluation through mandatory, directed, and additional comments. Specifically, the mandatory comments are intended to provide “a more complete and detailed evaluation of the MRO’s professional character and may address any entry made in sections A through H or as the Reporting Senior deems appropriate” and should address topics such as “performance, proficiency, potential, and other traits that describe the MRO utilizing the ‘whole Marine’ concept” (Commandant of the Marine Corps, 2015, p. 4-39). The PES manual articulates the responsibility of the RS or RO to provide “specific comments on potential for promotion and assignments to command, staff, and advanced schooling” (Commandant of the Marine Corps, 2015, p. 4-19).

Beyond that last bullet, the PES manual does not provide additional guidance on how to write the mandatory comments nor does it assist the RS in constructing statements that are consistent with the value of the MRO's performance. Particularly, MCO P1610.7 directs the RS to ensure consistency in reporting. While the Marine Corps Order does not instruct the RS to "'match' the attribute markings with the Section I comments," it does provides the guidance that the RS "must take care when making Section I comments to ensure that the comments neither conflict with, nor obscure, the remainder of the evaluation" (Commandant of the Marine Corps, 2015, p. 4-39).

The directed comments have prescribed structure and are required in the fitness report to provide the Commandant of the Marine Corps (CMC) amplifying information concerning the MRO. They fall into two categories: a venue to explain or enhance the 13 opportunities to mark fitness reports as commendatory or adverse in Section A, and 28 opportunities to augment the word picture made by the mandatory comments with observations that would draw attention to the promotion board on the promotability, retention, and assignment of the MRO. Common directed comments include:

1. Awards
2. Fitness reports with less than 90 days of observation time
3. Adverse fitness reports
4. Whether the MRO is filling a billet designated for a higher rank
5. Comment on flying proficiency
6. Extent to leaders apply operational risk management (ORM)
7. MRO's progress in professional development
8. Submission of an observed reporting when the reporting period is less than 90 days
9. Service in a combat zone

While not every fitness report has a required directed comment, our data set benefits from having leaders and aviators that all require directed comments on the MRO's compliance with ORM policy and the MRO aviators flight proficiency.

Finally, the RS has an opportunity to address a variety of events, accomplishments, and activities that are not directly linked to a performance attribute but contribute in building an overall picture of the MRO for the CMC (Commandant of the Marine Corps, 2015, p. 4-38).

Section J renders the document official by including the MRO, RS, and RO's signatures. If the report is adverse, the Marine has the option of explaining the adversity with an addendum page. Section K "formalizes the reviewing officer's involvement in the report" (Commandant of the Marine Corps, 2015, p. 4-46). The RO is charged with ensuring the administrative correctness of the report and to provide a comparative assessment of the MRO against all those reviewed within the same rank. In item 1 of section K, the RO indicates whether he or she has had sufficient observation and knowledge of the MRO during the reporting period to complete the assessment. In item 2, the RO indicates whether he or she agrees with the RS's evaluation of the Marine by selecting either 'concur' or 'do not concur'. If the RO does not concur with the assessment, he or she must provide remarks to amplify the disagreement; however, "non-concurrence is not considered adverse" (Commandant of the Marine Corps, 2015, p. 4-11) . In Item 3, the RO compares the MRO to others "of the same grade both past and present whose professional abilities are known to the RO" (Commandant of the Marine Corps, 2015, p. 4-47). This comparative assessment is distributed through a "Christmas tree" motif (Clemens et al., 2012) with the decreasing concentrations of observations as MROs ascend the distribution. An unsatisfactory marking by the RO renders the report adverse.

Even if the RO concurs with the RS's evaluation of the MRO, the PES manual does not require him or her to write a comment; however, it is common practice to do so. The manual guides the RO to consider information available such as official military records and comment on the MRO's performance during the reporting period, to "amplify his or her comparative assessment mark, and to evaluate the MRO's potential for continued professional development (e.g., promotion, command assignment, resident PME, and retention)" (Commandant of the Marine Corps, 2015, pp. 4-47 and 4-48). As appropriate, such as when the RS's profile is too sparse for meaningful value or confined

to a homogenous group of Marines, the RO ought to put the RS's marks and comments in perspective to inform readers on special circumstances attributed to the RS or MRO.

The reporting senior's relative value and reviewing officer's comparative assessment are some of the novel feature introduced in the A-PES (Phillips & Clemens, 2011, p. 4:5). A profile is a snapshot of the RS's and RO's "rating history, and includes information on the number of reports written, the fitness report averages for each grade, and the highest and lowest averages submitted by the RS and RO" (Clemens et al., 2012). They aid the Marine Corps in maintaining the integrity of the system and select the most qualified officers for retention and promotion. The relative value for each fitness report stems from the raw score provided by averaging the attributes in Section D through H and is calculated after an RS has accumulated no less than three fitness reports for a specific rank and uses the following equation:

$$\max\left(80, \frac{RawScore - RSaverage}{RS\max - \mathrm{RSaverage}} * 10 + 90\right).$$

The relative value is captured at the time of processing and continuously updated as the RS writes more fitness reports. These relative values are used to compare the MRO's performance against others written by the RS and are "displayed on the Master Brief Sheets of Marines and kept in their official military personnel files" (Ergun, 2003, p. 31).

The Marine Corps classifies Marine performance into thirds to account for minor variability in numeric assessments. Promotion, education, and other administrative boards have access to all FitReps that have not been administratively extracted from a Marine's official military profile. The electronic user interface which accesses the FitReps includes a briefing guide and summary statistics on relative values and comparative assessments. This display provides the number of FitReps an MRO "has received that fell in the upper third (RV > 93.33), middle third (86.66 < RV < 93.34), and lower third (RV < 86.67) of the reporting senior's profile. It also shows the number of RO assessments—from this officer's ROs to other MROs in the same grade— that were above, with, and below the mark they gave this officer" (Clemens et al., 2012, p. 55). Table 1 illustrates how relative

values and comparative assessments are displayed, reflecting the marks at the time of processing and cumulatively.

Table 1. Briefing Guide for a Fictional Officer. Source: Clemens et al. (2012, p. 55).

	At	Cum.
	RV Summary	
Upper third	2	2
Middle third	3	5
Lower third	2	3
N/A	4	1
	RO Assessment	
Above	10	24
With	19	35
Below	16	22

The reviewing officer profile provides an overall comparative assessment on an integer-scale from 1 to 8, with an intended distribution shaped like a "Christmas tree." This differs from the RV because "it is not derived from other numbers but is a directly assigned relative assessment" (Clemens et al., 2012, p. 8).

B. FITNESS REPORTING WRITING EDUCATION FOR OFFICERS

The only mandatory training for all reporting seniors and reviewing officers occurs during The Basic School. The Basic Officer Course manual concept card calls for presenting a one-hour lecture by the administrative leader and a two-hour workshop conducted by the staff platoon commander (The Basic School, 2016a). The one-hour lecture contains material on background information, the mechanics of running the A-PES interface, assigning marks, writing the word picture, the relative value, reviewing officer comments, and adverse reports (Dodd, 2016). The class does not provide information on how the relative value is calculated. The instructor staff recommends using the MCO P1610.7 and a list of the RVs previously written by the RS.

At the fitness report workshop participants write a fitness report as an assessment and discuss the results based on the their use of The Basic School's Fitness Report

Handbook (The Basic School, 2016b). The handbook provides some general guidelines and three examples of how to write a fitness report. The handbook references five sources used to create the period of instruction:

1. NAVMC 2794: How to Write a Fitness Report,

2. MCO P1070.12: Marine Corps Individual Records Administration Manual (IRAM),

3. MCO P1900.16 Marine Corps Separation and Retirement Manual (MARCORSEPMAN),

4. SECNAVINST 1650.1: Navy and Marine Corps Awards Manual and MCO P1610.7

5. Performance Evaluation System (PES) Manual (The Basic School, 2016). Students at The Basic School receive no further training on writing fitness reports.

Although NAVMC 2794 provides guidance on how to write FitRep comments, it is the old version of the PES manual and has been superseded by MCO P1610.7. The former provides direction on what and how to write Section I comments. It emphasizes that the narrative should be consistent with markings in Section B and directs the RS to provide an account of the MRO's successes and failures in performance, to discuss the MRO's potential to handle positions of increased responsibility, and to offer observations on skills and character. For the mechanics of writing, NAVMC 2794 specifies that the narrative start with a lead statement, amplify duty assignments, evaluate performance, provide insight on skills and character, and close by discussing potential of the MRO. For the structure of the comments, the manual directs the RS in several areas: use of simple factual statements, setting the tone in the first sentence, avoiding fillers and adjectives, condensing writing, eliminating MRO's name from the comments, use of bullets separated by semi colons, and starting each bullet with an active practical verb (USMC, 1995). Most importantly for the purposes of this thesis and of general fitness report writing, the manual provides a subject guide to address specific to rank, useful topics to address for promotability, and examples of fitness reports for different levels of performance. Although the old and new PES manuals are administratively similar

(Clemens et al., 2012), their contents are distinct with the latter providing little guidance on writing fitness reports. It is also noteworthy that the latter does not reference the former so the reader is not directed to material that would provide useful guidance for writing FitReps.

A key resource Marines utilize to guide them through the FitRep comment section is the commonly available “Fitness Report Writing Guide for Marines” written by Drewry (1998). Although not a sanctioned source, it is sold at nearly every Marine Corps installation book store. Written in 1986 with a last addition published in 1998, the book translates the NAVMC 2794 guidance into actionable information. It provides a general outline of an opening remark, comments on performance and character, and concludes with comments on promotion, retention, and assignment. Beyond providing structure, the book offers key phrases, power words, and examples for each tier of Marines. The use of this book is prevalent throughout the Marine Corps as evidenced by the structure and usage of key phrases. The book was written under the paradigm of the old fitness report system, however, which does not translate to the intent of the new PES.

Additional guidance is provided by the Marine Corps’s Expeditionary Warfare School and Command and Staff College, which conduct small group discussions on fitness reports; however, attendance at these schools is low and the non-resident programs do not address fitness reports.

In their 2012 review of the performance evaluation system, CNA identifies a considerable deficiency in training and education, which led to Manpower Management Division’s engagement in “developing training to help Marines understand how boards interpret and use FitReps” (Clemens et al., 2012, p. 53). Specifically, CNA observes that The Basic School does not explain how to generate RVs and gives an incorrect impression that the RV automatically normalizes fitness report averages (FRA) into a “bell curve” distribution (Clemens et al., 2012). The recommendations made by CNA had not been implemented as of the time this thesis was written.

C. LITERATURE REVIEW

Our literature review discusses previous work on fitness reports and general academic knowledge of sentiment analysis. Clemens et al. (2012) focuses on fitness report markings to express relationships between attributes of the MRO, RS, and RO, and categorical variables that include demographics and various indicators of performance. Jobst and Palmer (2005) examine fitness reports to determine whether entry-level performance and commissioning source are useful indicators of current performance. For text mining, we examine literature to identify text-based features and techniques for sentiment classification based on product reviews, trip reports, and university professor evaluations. To our knowledge, our thesis is the first analysis that considers the textual information in fitness reports.

1. Study by The Center for Naval Analyses

The Center for Naval Analyses (CAN) 2012 study was commissioned by the "Director, Manpower Management Division (MM) to determine whether the performance evaluation system is accomplishing what the Corps intended. She requested that CNA "focus on officers and consider whether the new system is keeping inflation in check, ensuring fairness for all officers, and helping the various boards select the 'best and most qualified' officers" (Clemens et al., 2012, p. 1). The study provides a comprehensive overview of A-PES by examining the marks, biases based on general observable characteristics such as race and gender, and how scores are presented to the boards.

Lacking the physical and computational ability to read all comments, CNA (2012) reviews a small sample of 300 fitness reports of only captains with specific racial backgrounds in evenly spaced tier groups from an RV of 80 to 100 in increments of four points (Clemens et al., 2012, p. 47). Based on a reading of each report, the authors focus only on the recommendations for promotion through the classification table provided in Table 2. Clemens et al. (2012) conclude that when comparing equivalent marks between racial and gender groups, the gap in observed markings is not statistically significant.

Table 2. Subjective Comments Classified into tiers of Promotion Recommendation Strength. Source: Clemens et al. (2012, p. 48).

Tier 4	Tier 3	Tier 2	Tier 1
"I do not recommend promotion"	"promote"	"enthusiastically recommended for promotion"	"promote ahead of peers"
"qualified for promotion"	"promote with peers"	"promote at first opportunity"	"groom for highest ranks in Marine Corps"
(nothing)	(implied by recommendation for battalion command)	"highly recommended for promotion"	"my highest recommendation for promotion"
		"promote now"	"a must for promotion"

The CNA's analysts at the Marine Corps' Operations Analysis Division (OAD) recognize some "potentially confusing differences between how RVs and RO marks are tabulated" (Clemens et al., 2012, p. 56). For example, "larger numbers in the top row of the RV summary are good" for the MRO, "whereas larger numbers in the bottom row of the RO assessment are good" (Clemens et al., 2012, p. 56). The CNA study concludes that the FitRep system is working well but could improve on the training and education of FitRep writers. Further, CNA recommends that the ROs state their level of familiarity with the MRO (Clemens et al., 2012).

2. Study by Mark Jobst and Jeffrey Palmer

Jobst and Palmer (2005) study the FitReps as part of a joint thesis to fulfill the obligations for a Master's of Science in Management at the Naval Postgraduate School. The authors' research addresses three topics: "Firstly, …provide validity for the two-sided matching process; secondly, analyze FitRep attributes to determine their suitability for a weighted criteria evaluation system and; thirdly, compare the USMC promotion and assignment process with contemporary human resource management practices" (Jobst & Palmer, 2005, p. v). Their analysis on fitness reports is a component of an overall better human resource management outline for the Marine Corps

The authors conclude that not all marks in Sections D through H are equally weighted with top relative value performers having proficiency and performance marks higher than the other 12 (Jobst & Palmer, 2005, p. 73). The relative values are biased in multiple ways:

- “Central Tendency – everyone is rated in the middle” (Chmiel, 2000, p. 131);
- “Halo Effect – assessment of one quality of the individual affects the judgment of all his or her other attributes, so all ratings are highly correlated” (Chmiel, 2000, p. 131);
- “Positive Skew – everyone is rated high (all swans, no geese)” (Chmiel, 2000, p. 131);
- “Recency Bias –because managers rarely keep detailed notes about their appraises, and are not very precise about rating all the behaviors they are required to judge, there is a tendency to base appraisal on the recent past, regardless of how representative it is of performance over the year” (Bach & Sisson, 2000, p. 252)

Additional biases that we expect to encounter in the study are rank bias, where the higher the rank, the higher the value; and communities bias, where communities, such as combat arms, aviation, aviation support, or combat service support, tend to engage in self-preservation behavior when mixed with other groups.

Similar to the CNA study, Jobst and Palmer recommend more education and training for RS and RO “such as rater error training, performance dimension training, frame of reference training, and behavioral observation training” (Jobst & Palmer, 2005, p. 71; Chmiel, 2000).

3. Dissertation by Donald Jordan

In his Doctoral dissertation, Jordan (2011) examines student course evaluations to extract sentiment. Course evaluations are widely used by educational institutions to provide feedback to instructors regarding their teaching. All these evaluations are composed in three basic forms: “1) a variety of statistical questions using multiple choice and Likert scale responses, 2) open-ended questions that allow students to respond with their own words, and 3) a combination of both” (Jordan, 2011, p. 1).

The student surveys considered by Jordan (2011) share similarities with fitness reports: both use Likert Scales to elicit levels of agreement or disagreement on a series of statements. Both also elicit comments in open-ended paragraphs to provide feedback. These comment fields fulfill the same purpose as Section I as they are a “catch-all for

students to write out their observations, recommendations, frustrations, and any other issues that may not have been addressed" (Jordan, 2011, p. 4).

Jordan (2011) is one of the first of its kind due to the computational difficulty in quantitatively analyzing textual. The study pulls 835 anonymous and non-attributable surveys between 2005 and 2009 from the Center for Professional and Continuing Education at the University of the Pacific in Stockton, California, and attempts to answer the following four questions:

1. Are the student comments of course evaluations aligned to the quantitative portion of the course evaluation instrument?
2. Are there words and patterns prevalent in the unstructured data of student comments of course evaluations that can classify individual courses on the basis of negative connotations?
3. Are there words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can provide additional information at a program or institutional level?
4. Is there an association between the results of the text mining analysis of the unstructured data and a qualitative analysis of the unstructured data? (Jordan, 2011, pp. 10-11)

Displayed in Figure 1, Jordan uses a variety of traditional text-mining techniques and models with Principal Component Analysis, Singular Value Decomposition, and K-Means clustering to answer his research questions. The text processing includes removal of stop words, punctuation, capitalization, stemming, and correcting for sparsity.

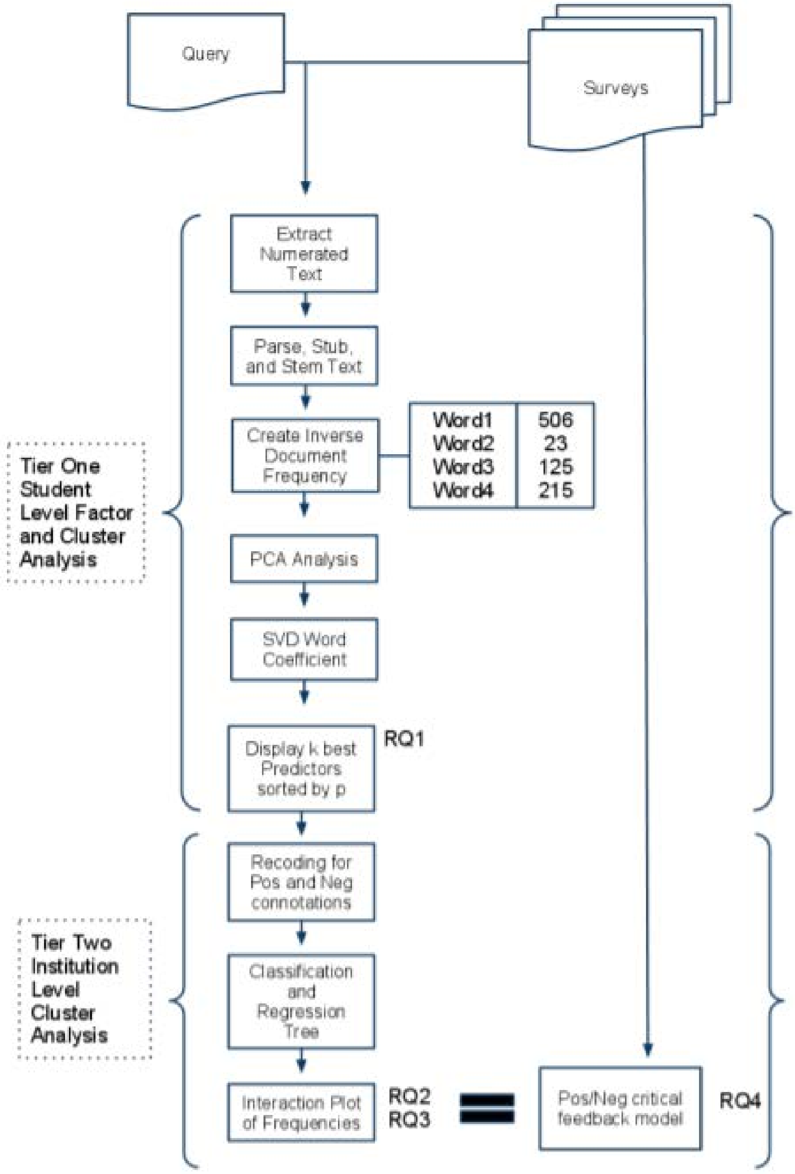


Figure 1. Revised Text Mining Workflow. Source: Jordan (2011, p. 49).

Once a matrix of documents and words is created and the modeling tools are applied, the author uses the results to classify the comments as either positive or negative feedback through sentiment analysis.

The study's major finding show is that while there is only a "weak correlation between the Likert responses and the open-ended written portion, there are significant words and patterns within the unstructured data that provide additional information at the

institutional level” (Jordan, 2011, p. ix). Validated through K-Means clustering, Jordan yielded two single word lexicons that correlate to positive or negative sentiment. The author suggests that due to the lack of structure and poor correlation between marks and comments, “colleges need to rethink the design, implementation, and approach to the student course survey that can take advantage of text mining as an analytical tool for the institution” (Jordan, 2011, p. ix). Recommendations include implementing standardized submittal, incentives participants, removal of anonymity, and add structure to make it text mining friendly.

4. Study by James Friedlein

In partial fulfillment of his Master’s degree in Operations Research from the Naval Postgraduate School, Friedlein (2016) implements a cascade classification model on Islamic State (IS) press releases. Using 2,926 IS press releases collected by the NPS Defense Analysis Professor Craig Whiteside, Friedlein considers data from multiple terrorist databases and classifies the press release subject according to global terrorist incidents. He uses a text processing approach to that is similar to Jordan (2011) and considers two additional models: a regularized generalized linear model and a cascade classifier, which discards background information to focus on promising information (Friedlein, 2016). After cross-validation, the models produces a misclassification rate of 5.5 percent, rendering these models as worthwhile approaches for text classification (Friedlein, 2016, p. 32).

5. Study by Anindya Ghose and Panagiotis G. Ipeirotis

Ghose and Ipeirotis (2011) study pre-corpus linguistics statistics embedded in online product reviews. They examine “the impact of reviews on economic outcomes like product sales and see how different factors affect social outcomes such as their perceived usefulness” (Ghose & Ipeirotis, 2011, p. 1498). Specifically, they explore “subjectivity levels, various measures of readability, and extent of spelling errors to identify important text-based features” (Ghose & Ipeirotis, 2011, p. 1498) without looking into tokenized word models. Ghose and Ipeirotis used a random forest classification model because they are robust and perform better than support vector machines for a variety of learning tasks

(Ghose & Ipeirotis, 2011, p. 1508). The authors looked at the review's *helpfulness* as related subjectivity, spelling, and readability. The conclusions are presented in the area under the ROC curve and show that subjectivity and readability are the most valuable factors.

Their results led to some interesting observations. Readability statistics such as Automated Readability Index (ARI), Coleman-Liau index, Flesch-Reading Ease, Flesch-Kincaid Grade Level, and the Simple Measure of Gobbledygook (SMOG) index are helpful in predicting positive reviews and associated with higher sales (Ghose & Ipeirotis, 2011, p. 1505). Conversely, the presence of spelling errors have a statistically significant negative impact on the helpfulness of a review and yielding to lower sales (Ghose & Ipeirotis, 2011, p. 1511).

6. Study by Kushal Dave, Steve Lawrence, and David Pennock

Dave, Lawrence, Pennock's (2003) research contribute to the development of an opinion mining tool that "process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)" (Dave, Lawrence, & Pennock, 2003, p. 519). The authors develop a tool that "synthesized product reviews, automating the sort of work done by aggregation sites or clipping services" (Dave, Lawrence, & Pennock, 2003, p. 519). Figure 2 displays their work flow of this tool. They begin by using structured reviews for "testing and training, identifying appropriate features and scoring methods from information retrieval for determining whether reviews are positive or negative" (Dave, Lawrence, & Pennock, 2003, p. 519).

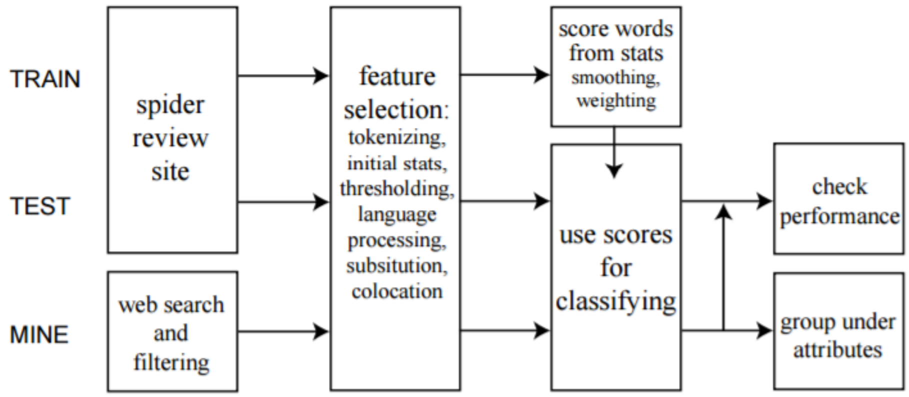


Figure 2. Overview of Project Architecture and flow. Source: Dave, Lawrence, and Pennock (2003, p. 1).

Their research is similar to aforementioned research but is distinct in two respects. First they depart from single word analysis by combining words together into tokens and they add groups of words that immediately precede the product name or come shortly after. This technique yields powerful results that lead to substantially better classification rates (Dave, Lawrence, & Pennock, 2003, p. 522). Second, they do not remove punctuation. Contrary to other text mining research, their analysis determines that reviewers put their strongest impressions at the beginning or end of a sentence to highlight the sentiment. Proximity to the period yields better classification rates of positive or negative sentiment (Dave, Lawrence, & Pennock, 2003, p. 525).

The authors also develop a classification algorithm that works better than traditional machine learning techniques. Their results, however, exhibit high variability due to the relatively small size of the sample, and the authors acknowledge that they must deal with heavy overfitting. Increasing the sample size and adding granularity to the review tags would increase the accuracy of their results.

III. DATA AND METHODOLOGY

This chapter provides extensive detail on how we handle the data and the methodology we use in our analysis. We take our data from two different sources and convert them to a single, cleaned data set that we can use in modeling. The text fields are processed and simplified to account for slight wording-variations while preserving the meaning and intent of the words. In preparation for modeling, we generate specific, specialized matrices to capture frequency of words in each document within the data set.

The second part of the chapter summarizes the techniques we implement in our statistical analysis and predictive modeling. These methods include readability metrics, non-parametric statistics, and supervised machine learning algorithms. To facilitate a three-tiered classification of reviewing officer comparative assessments, we develop a technique to classify a MRO into three performance tiers.

A. DATA DESCRIPTION

The data collected for this thesis were obtained from the PES database and A-PES application, which is maintained by Performance and Evaluation Section, Manpower Management Records and Performance Branch (MMRP), Marine Corps Manpower and Reserve Affairs (M&RA) in Quantico, VA. In January 1999, the Marine Corps adopted the Performance Evaluation System (PES); however, the database only stores the administrative section and raw scores. Text blocks (sections I and K) are not stored in the PES database; only PES data needed to populate the Master Brief Sheet and index fitness reports in the Official Military Personnel File (OMPF) is stored. The raw data is used to calculate relative value and other numeric metrics are computed to categorize the performance of the MRO.

Prior to 2006, sections I and K comments were not stored in the PES database because: 1) The Optical Character Recognition (OCR) process that was used to read the text fields of scanned, paper reports was not accurate; therefore, requiring the re-keying of text on the majority of the form; 2) MMRP did not have the resources to edit all the comment blocks on scanned, paper reports; 3) Only the image and specific data from the

form was needed to populate a Master Brief Sheet (Marucci, 2016). Before A-PES, reporting officials submitted fitness reports via WinFE or FormFlow. These systems semi-automated the creation and workflow of the form but the forms were still printed and scanned into the PES Back Office system for editing, correction, and processing. The scanned forms where then run through OCR software that attempted to read the text on the form. Since the OCR was not very accurate, it was decided that only fields used to populate the Master Brief Sheet would be edited and stored in the PES database.

In 2005, the A-PES application was adopted to help reporting officials electronically create and route fitness reports up the reporting chain and to submit them to MMRP. Although the A-PES system is not the system of record and was not developed for that purpose, all fields, including sections I and K, are stored in the A-PES database for all reports submitted through that system. Currently, A-PES is used almost exclusively to submit reports, although, PDF forms can still be used to submit fitness reports when circumstances prevent the use of A-PES. As fitness reports transitioned to online submissions, they began to populate the A-PES database; however, the data in A-PES is only the data submitted to PES (e.g. A-PES data will not change if a change is made to the report after submission.) MMRP has authority to make administrative corrections to reports after submission, which usually do not include text blocks (Section I and K). If a correction to a text block is identified before processing, the report is returned to A-PES for that correction. Fitness Report corrections after processing that are performed by MMRP for administrative reasons or by an approved Performance Evaluation Review Board (PERB) case may not be reflected in A-PES. However, if PERB approves the pulling of an entire report from the system, it is also removed from A-PES. In 2006, 70% of the fitness reports' Sections I and K comments written that year were in the system; 2007, 97%; and afterwards over 99% with a few outliers in non-traditional Marine Corps commands (Marucci, 2016).

Because of the lack of availability of comments prior to 2006, the data does not include records prior to 2006. With data only available between 2006 and 2016, a complete longitudinal study from entry to retirement is not possible; however, a cross-sectional analysis of each rank during this time period may be conducted. To capture the

company grade officer slice, the 2006 and 2007 cohorts of officers are followed from entry to 2016. This 10 to 11 year slice has two years of observable time for second and first lieutenants and six years for captains for the officers that follow traditional promotion rates. To capture the field grade stratum, the 1996 and 1997 cohorts of officers are followed from 2006 to 2016. This 11-year slice includes 6 years of observable time for majors and four years of lieutenant colonels at traditional promotion rates. These slices represent 5,596 officers of whom 4,761 have observed fitness reports in the system. Displayed in Figure 3 is the quantity breakdown by rank of observed reports we considered for analysis. The Marine Corps does not accelerate the promotion rates of officers beyond the cohort year group; however, officers can fall behind. Officers fail to be selected for promotion for a variety of reasons such as observed performance metrics falling below expectations, failure to complete required grade education, or other issues outlined in the PES manual.

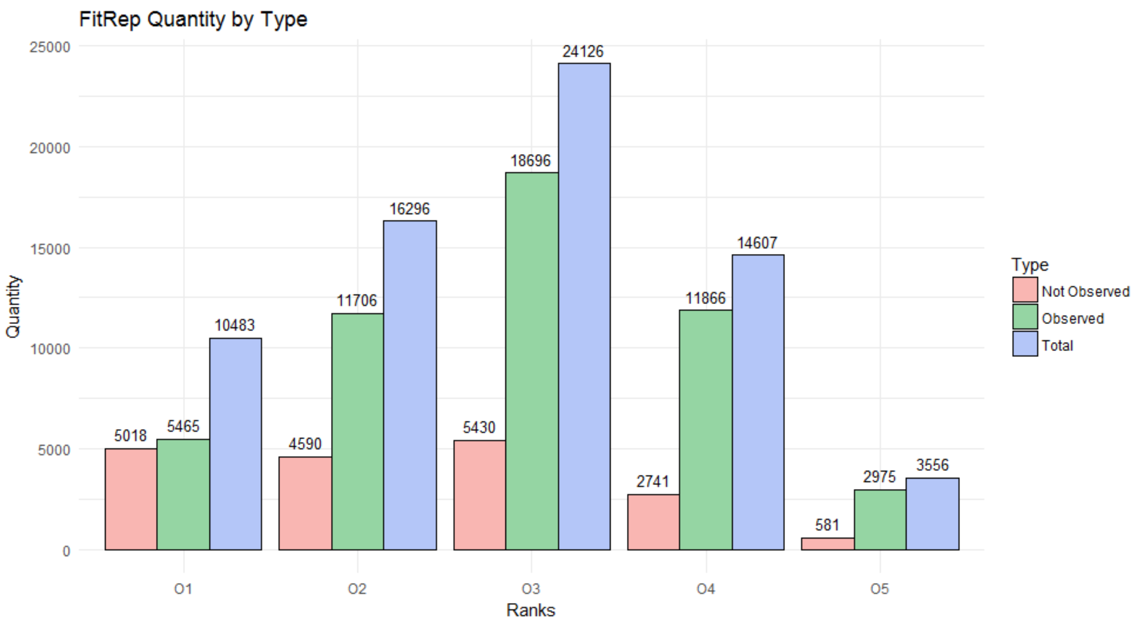


Figure 3. Rank Breakdown of Observed to Not Observed FitRep Quantities

The data from the “PES Back Office” is a spreadsheet that is a 71,212 × 81 matrix. Each row constitutes a single each fitness report. The columns represent the administrative fields, the attribute markings, the RS relative values, and RO comparative

assessment mark, which are outlined in Appendix B, Table 23. The administrative section places a person, location, time, annual required qualifications, into temporal context within the field. Case numbers are used in lieu of names or Department of Defense identification numbers to preserve the identity of the Marine, which is not required in the analysis.

The next section identifies the fields that are evaluated on in each fitness reports. The 14 attributes are scaled from “A,” which is adverse and numerically represented by 1, to “G,” which is outstanding and represented by 7. An “H” is marked when the RS does not observe a specific MRO trait during that reporting period. When the average of all traits is calculated and “H” is selected, the denominator is reduced by the numbers of “H”s.

The core of the first data base are the MRO’s assessment values based on the RS and RO profiles. Each of the fields under the RS assists in populating the relative value and place the MRO’s performance into context. Concurrently, the RO fields help populate the RO’s comparative assessment markings; however, the RO does not have a relative value. Marucci (2016) propose the following formulas to quantify MRO performance with respect to the RO where the ROV is the weighted average of the comparative assessments

$$ROV = \frac{\sum_{i=1}^{8} ROAssessment_i \times AssessmentCount_i}{\sum_{i=1}^{8} AssessmentCount_i},$$

where the RO Assessments is the sum To classify the performance of the MRO as above or below average, the difference between the RO raw score and the weighted average is computed with

$$ROVDiff = RawScore - ROV \quad .$$

The text data from “A-PES” correspond to section I, section K, and Addendum page comments. Each row represents an individual report and is merged to the “PES

Back Office" data through a case ID and the report's submission date. We bind all the fields together to commence text processing.

B. TEXT PROCESSING

During this section, we take the data set and we prepare for analysis. We start by separating training, validation, and testing sets to use in our predictive models. We process character strings into usable form by removing excess entities such as special characters, numbers, and punctuation. We use a bag-of-words approach which is a natural language processing technique that takes *n-grams* and tokenizes them into a single object. The term *gram* can be viewed as a contiguous sequence of *n* items represented in a source text. Grams go beyond words as they can capture syllabus, single letters, or numbers. Each combination of *n*-grams becomes a token that we use to track frequency across each FitRep throughout the *corpus.* A corpus is the collection of structured documents. We fit the corpus into a matrix with each row representing a FitRep and each column representing a token. These document-term matrices are the root structure of the data we use for subsequent modeling.

1. Separation of Training, Validation, and Test Sets

Prior to analysis, the data are separated into a training set, a validation set, and a test set. During this study, the training set is used to fit the each of the models. We estimate the prediction error for each model selection through the validation set. Finally, once we select the final ensemble of models, we assess the generalization effort using the test set, which is locked into a "vault" and only brought out only at the end of the analysis (Hastie, Tibshirani, & Friedman, The Elements of Statistical Machine Learning: Data Mining, Inference, and Prediction, 2009, pp. 222-223).

Because of we have a data rich set of 71,212 observations, we are able to afford a 25% validation set and 10% test set. Even though we are in a data rich environment the data become sparse over sub groups such as commendatory and adverse level FitReps. Furthermore, we know the hierarchal nature of the military ensures that there are more junior officers than senior officers. As a result, we stratify the separate training, validation, and test sets by taking the same proportion of FitReps in each category

outlined below, which will ensure the different strata contain enough representation to build and evaluate a classification model.

- By report type
  - Adverse
  - Normal
  - Commendatory
- By tiers
  - Bottom Third (RV < 86.67))
  - Middle Third (86.66 < RV < 93.34)
  - Top Third (RV > 93.33)
- By Rank
  - Second Lieutenant (O1)
  - First lieutenant (O2)
  - Captain (O3)
  - Major (O4)
  - Lieutenant Colonel (O5)

2. Text Processing

We process the data with the work flow outlined in Figure 4. We clean remove fitness reports that do not provide information such as unobserved reports and FitReps tied to low-density profiles. We process the corpus of fitness reports into n-gram tokens to the term document matrices.

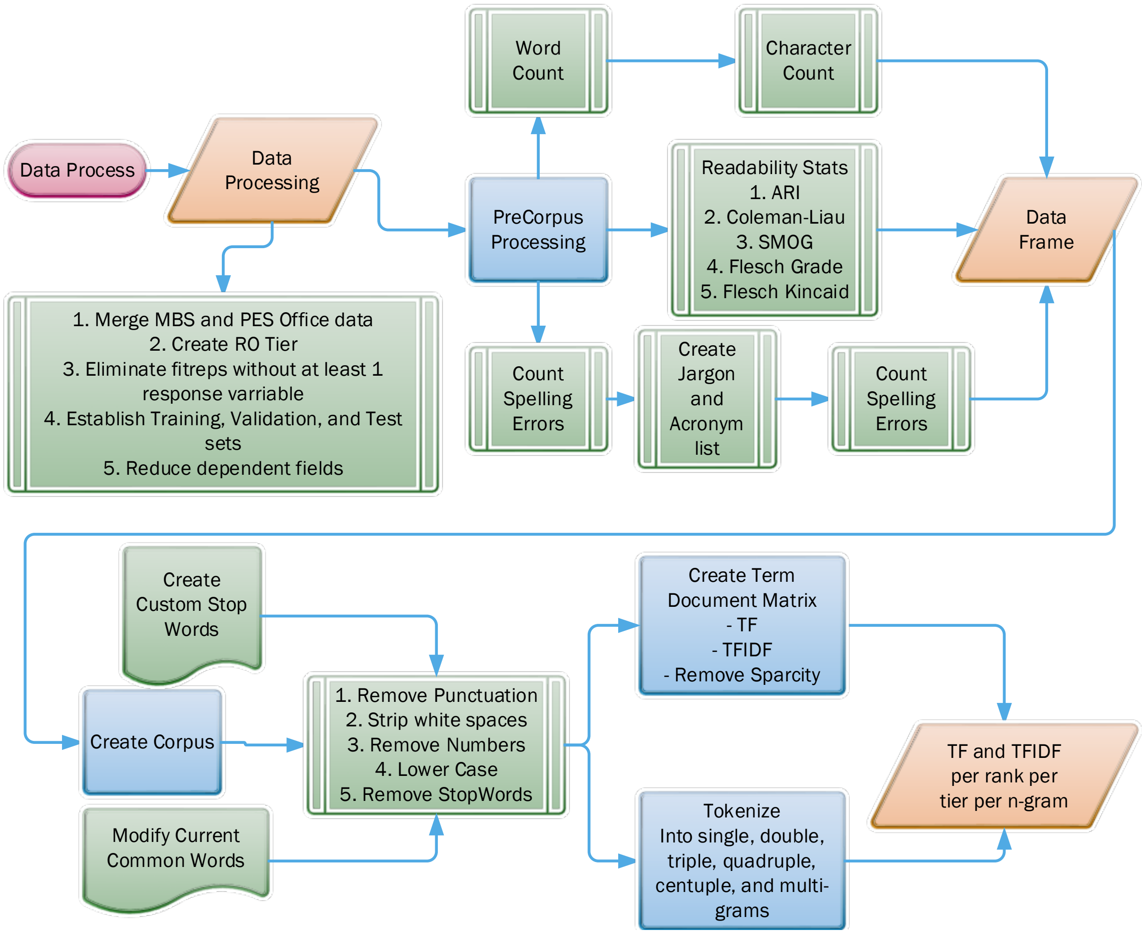


Figure 4. Data Processing Process Flow

The 71,212 fitness reports have an average of 731 characters in section I and 389 characters in section K comments, which requires special consideration due to the large volume of information. Using the **tm** package of R (Feinerr & Hornik, 2015), we create the corpus to handle analyzing the text and perform other natural language processing functions. We use the volatile corpus (VCorpus) to store the file in memory on a local machine. Each rank and tier is divided and incorporated into a corpus. **VCorpus** takes a data set of character data and constructs a vector with a name corresponding to each FitRep to contain the textual information and file metadata.

The text contained in the corpus is considered a "bag of words" and is treated as an unordered collection of words or tokens (Friedlein, 2016, p. 13). The corpus requires basic text mining operations that are provided by the tm_map() function in the **tm**

package. The common operations are removal of punctuations, numbers, capitalizations, extra white space, special characters, stop words, and word-stemming. Stop words are common words such as “and,” “the,” “how” that do not provide additional value in textual analysis and have the potential to affect our findings due to their common frequency. Word-stemming is the process of reducing a word to its root form by removing prefixes and suffixes. While we lose some of the context of the word use, we retain its meaning and can account for small variations of its use in a sentence. To illustrate the preprocessing steps, a section I comment is selected and demonstrated below:

Original text

“#1 captain in the battalion. MRO is one of the most talented and gifted minds we have in the Marine Corps. He is a strategic thinker, a systemic planner, a superb tactician and the most physically fit Marine I have ever served with in the Corps, which again speaks to why his was selected as a Leftwich Award finalist by the Corps. In the period of time he has served as the operations officer, MRO has made a tremendous and lasting impact on this battalion through his detailed planning, exhaustive coordination and flawless execution of the battalion’s deployments and execution of Mountain Exercise 215, Exercise Balikatan 15 and Lava Viper 15. This talented officer is wise beyond his years and can compete pound for pound with any seasoned major as a battalion operations officer. Already selected for promotion to major and ILS, continue to promote this officer, place in billets where the Corps needs are best and brightest. Finally, there is no doubt in my mind, this officer’s will and should be command slated, in the future, as an active component infantry battalion commander. Directed Comments, Sect A, Item.6a: During this reporting period, MRO was selected as a 2015 Leftwich Award finalist and awarded the Navy/Marine Corps Commendation Medal for superior sustained performance of his duties during his tour in the battalion.”

Removal of numbers and punctuation

“captain in the battalion MRO is one of the most talented and gifted minds we have in the Marine Corps He is a strategic thinker a systemic planner a superb tactician and the most physically fit Marine I have ever served with in the Corps which again speaks to why his was selected as a Leftwich Award finalist by the Corps In the period of time he has served as the operations officer MRO has made a tremendous and lasting impact on this battalion through his detailed planning exhaustive coordination and flawless execution of the battalion’s deployments and execution of

> Mountain Exercise Exercise Balikatan and Lava Viper This talented officer is wise beyond his years and can compete pound for pound with any seasoned major as a battalion operations officer Already selected for promotion to major and ILS continue to promote this officer place in billets where the Corps needs are best and brightest Finally, there is no doubt in my mind this officers will and should be command slated in the future as an active component infantry battalion commander Directed Comments Sect A Itema During this reporting period MRO was selected as a Leftwich Award finalist and awarded the NavyMarine Corps Commendation Medal for superior sustained performance of his duties during his tour in the battalion”

There is some information lost by cutting the numbers. Many reporting seniors like to place a relative rank of the MRO relative to his peer group because the old performance evaluation system encouraged it (USMC, 1995, pp. 6–7), or the reviewing officer is guided towards commenting on his comparative assessment (Commandant of the Marine Corps, 2015, p. 4-48); however, this requirement no longer is necessary (Commandant of the Marine Corps, 2015, pp. 4–38:4-40). The current A-PES calculates relative and comparative assessments at the time of processing and cumulatively. As a result, removing a relative ranking when the data are already divided into tiers may eliminate information on whether the officer is at the top or the bottom of the third, but does not influence the classification of the fitness report.

Change to lower case

> “captain in the battalion mro is one of the most talented and gifted minds we have in the marine corps he is a strategic thinker a systemic planner a superb tactician and the most physically fit marine i have ever served with in the corps which again speaks to why his was selected as a leftwich award finalist by the corps in the period of time he has served as the operations officer mro has made a tremendous and lasting impact on this battalion through his detailed planning exhaustive coordination and flawless execution of the battalion’s deployments and execution of mountain exercise exercise balikatan and lava viper this talented officer is wise beyond his years and can compete pound for pound with any seasoned major as a battalion operations officer already selected for promotion to major and ils continue to promote this officer place in billets where the corps needs are best and brightest finally there is no doubt in my mind this officers will and should be command slated in the future as an active component infantry battalion commander directed comments sect a itema during this reporting period mro was selected as a leftwich award finalist and awarded the navymarine corps commendation medal for

superior sustained performance of his duties during his tour in the battalion"

The **tm** package uses the "Porter Stemming Algorithm for term normalization in information retrieval systems" (Feinerr & Hornik, 2015). Stemming of word involves removing the suffix from the root word to account for words used in slightly different context but retain the same meaning.

> "captain in the battalion mro is one of the most talent and gift mind we have in the marin corp he is a strateg thinker a system planner a superb tactician and the most physic fit marin i have ever serv with in the corp which again speak to whi his was select as a leftwich award finalist by the corp in the period of time he has serv as the oper offic mro has made a tremend and last impact on this battalion through his detail plan exhaust coordin and flawless execut of the battalion' deploy and execut of mountain exercis exercis balikatan and lava viper this talent offic is wise beyond his year and can compet pound for pound with ani season major as a battalion oper offic alreadi select for promot to major and il continu to promot this offic place in billet where the corp need are best and brightest final there is no doubt in my mind this offic will and should be command slate in the futur as an activ compon infantri battalion command direct comment sect a itema dure this report period mro was select as a leftwich award finalist and award the navymarin corp commend medal for superior sustain perform of his duti dure his tour in the battalion"

There are a multitude of stop-word dictionaries available through **R** that are compatible with the **tm** package; however, they remove potentially descriptive adjectives placed before performance words. As a result, we modify the available dictionaries by removing adjectives or qualitative words. Furthermore, we create a new dictionary of administrative words such as "directed comment" or "continue on addendum page" that are part of the FitRep syntax, but offer no value in quantifying the value of the words.

After removing the stop-words and stripping excess white spaces, we are left with.

> "captain in the battalion mro is one of the most talent and gift mind we have in the marin corp he is a strateg thinker a system planner a superb tactician and the most physic fit marin i have ever serv with in the corp which again speak to whi his was select as a leftwich award finalist by the corp in the period of time he has serv as the oper offic mro has made a tremend and last impact on this battalion through his detail plan exhaust

coordin and flawless execut of the battalion' deploy and execut of mountain exercis exercis balikatan and lava viper this talent offic is wise beyond his year and can compet pound for pound with ani season major as a battalion oper offic alreadi select for promot to major and il continu to promot this offic place in billet where the corp need are best and brightest final there is no doubt in my mind this offic will and should be command slate in the futur as an activ compon infantri battalion command direct comment sect a itema dure this report period mro was select as a leftwich award finalist and award the navymarin corp commend medal for superior sustain perform of his duti dure his tour in the battalion"

3. Creation of Reviewing Officer Tiers

As described in Chapter II, Section A, the reviewing officer assigns a comparative assessment score to the MRO. Unlike the reporting seniors' relative values, the comparative assessment are not evenly into even tiers during the summary process (Marucci, 2016). During promotion board reviews, the comparative assessment are categorized into three groups: those who were marked above, with, or below the MRO. For example, a marine would be below average if two thirds of the RO's marks were either above or with him. Prior research (Marucci, 2016) developed an equation to separate top and bottom half by providing a numerical distance from the mean based on a weighted average shown in Chapter II, Section III.A. This technique however, does not separate the comparative assessments into thirds. To create a three-tiered response variable for analysis, we design a tiered system outlined in Figure 5 that starts by counting how many officers are above, with, and below the MRO. Then, we isolate the middle and work through conditions to separate the top and bottom tiers. This results in fitness reports being distributed uniformly amongst the three tiers (up to rounding) at the conclusion of this process.

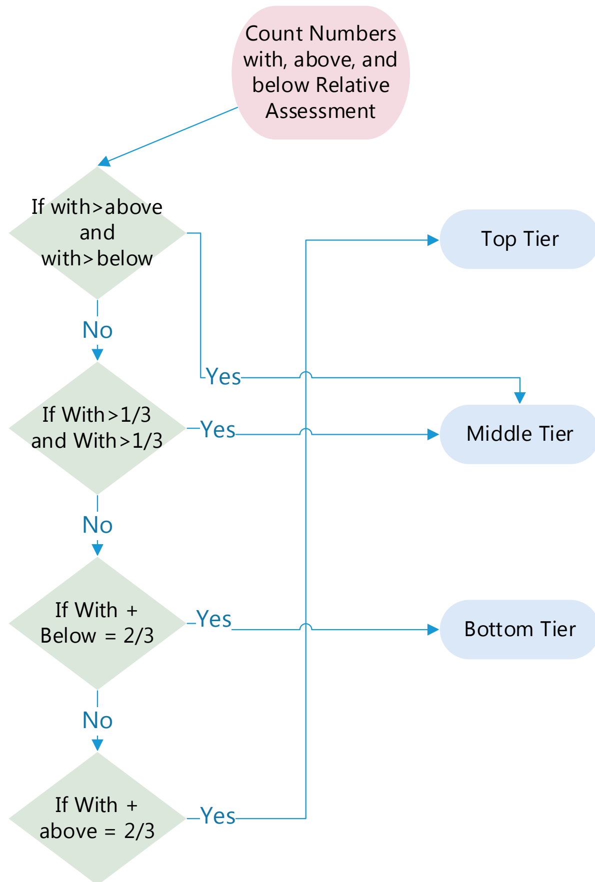


Figure 5. Process Flow for the Creation of Comparative Assessment Tiers

4. Creation of Document Term Matrix

a. Document term matrix

Once the character vector is processed through the **tm** wrappers, we store the frequency of terms in a document-term matrix (DTM) and term frequency-inverse document frequency matrix (TF-IDF). Through the process, the terms are transformed into tokens that represent either a word or a series of n-words. Taking an approach similar to Jordan (2011) and Friedlein (2010), we set up the corpus into matrices that feed into our predictive models.

b. N-gram Tokenizer

Using the simple DTM isolates tokens from their context and discards potential relationships between words. To avert this loss of information, we use consecutive n-gram tokens to preserve the relationships between words and locations in the sentence (Dave, Lawrence, & Pennock, 2003). Previous research using multiple n-gram tokens shows that using up to three-words has the most predictive power (Hulth, 2003). Though inclusion of longer expressions permit more expressive phrases, “systems that permit longer phrases can suffer from poor precision and meaningless terms” (Chuang, Manning, & Heer, 2012, p. 4) . There is a danger in including longer tokens as it may result in the “inclusion of longer phrases may also result in redundant terms of varied specificity such as ‘visualization,’ ‘data visualization,’ and ‘interactive data visualization’” (Evans, Klavans, & Wacholder, 2000). We will negate this tendency by using penalty-inducing models such as the lasso generalized linear model or algorithms that take into account interactions such as classification and regression trees.

The **RWeka** package (Hornick, Buchta, & Zeileis, 2009) provides a customizable function **NGramTokenizer()** that enables us to create the n-gram tokens based on a minimum and maximum size. To capture differences in relationships that may be drowned out by other tokens’ frequencies, we use tokens that increase in complexity starting with a unigram, or single-word token up to hexagram, or six-word tokens. We stop at hexagram frequencies based on the length of recommended promotion clauses by The Basic School (The Basic School, 2016c and Clemens, et al., 2012). Each DTM is run through the family of predictive models to select the model with the best F-Score rate.

c. Term Frequency-Inverse Document Frequency Weighting

The Term Frequency-Inverse Document Frequency (TF-IDF) is a “set of weights used to measure the significance of a term in a document” (Weiss, Indurkhya, & Zhang, 2010). In this design, terms that frequently occur in a document are given more weight, while terms that occur in most documents and have little predictive power (e.g., “Marine”) carry less weight. The common term frequency matrix normalizes the term by

$x_{ij} / \sum_j x_{ij}$ where x_{ij} is how often token *i* appears in document *j* (Friedlein, 2016, pp. 16-17). The inverse document frequency calculation provides a "dampening of simple word frequencies by using the log[arithm] function and includes a weighting factor *I* that evaluates to 0 if the word occurs in all documents *D* and assigns a maximum value if the word only occurs in one document" (Jordan, 2011, p. 54).

We transform taking the minus base-2 logarithm of the proportion of times that token *i* appears in a document

$$IDF_i = \log_2\left(D / \sum_{j=1}^{D} I(x_{ij} > 0) \right)$$

The process results in a transformation of the data with the creation of indices that reflect the relative frequency of word occurrence and semantic value within the documents included.

The final matrix is determined through

$$TFIDF_{ij} = (x_{ij} / \sum_j x_{ij}) \times IDF_i ,$$

which is calculated using the function weightTfIDF() in **tm** (Weiss, Indurkhya, & Zhang, 2010).

To determine which model configuration has the most predictive power we create 360 term document matrices: three tiers of five ranks and we examine n-gram token sizes from one to six for both the RS and RO. Finally, we look at the difference between term frequency-inverse document frequency and standard weighting. To ensure that we account for the variables and their interactions in the *n*-gram term document matrix, we retain all of the variables leading up to *n*. As a result, these matrices greatly increase in size, up to 25.2 gigabytes for the hexagram matrix.

To handle the size and number of matrices along with having to run multiple high-memory statistical machine learning algorithms, we use the high performance computer (HPC) at the Naval Postgraduate School. The Hamming supercomputer is a

"hybrid cluster" with 3,178 cores that possess different hardware specifications to solve academic and research problems (Naval Postgraduate School, 2017).

To run all these models in parallel, we modify an algorithm developed by Mckechnie (2017). We setup a series of parallel processes that create the 360 matrices, run them through various model configurations, and capture model performance characteristics.

C. STATISTICAL ANALYSIS

1. Readability Statistics

Ghose and Ipeirotis (2011) observe that the readability statistics of product reviews are a principal predictor of classifying positive and negative sentiment out of Amazon product reviews. Ghose and Ipeirotis use the "Automated Readability Index, Coleman-Liau index, Flesch-Reading Ease, Flesch-Kincaid Grade Level, and the Simple Measure of Gobbledygook (SMOG) index" (Ghose & Ipeirotis, 2011, p. 1505). Born out of the necessity to simplify the technical manual for enlisted personnel operating machinery, the popular Flesch-Kincaid readability statistics assign a quantifiable score to a document for refinement (Kincaid, Fishburne, Rogers, & Chissom, 1975, p. 1). Beginning with the Flesch-Reading-Ease test, scores a sentence by

$$206.835 - 1.015\left(\frac{totalWords}{totalSentences}\right) - 84.6\left(\frac{totalSyllables}{totalWords}\right)$$

(Kincaid, Fishburne, Rogers, & Chissom, 1975, p. 14).

According to Kincaid et al. (1975), a higher score indicates material that is simpler and easier to read while a lower number designates passages that are more difficult to read. The maximum achievable score is 120 and is a sentence of two single syllable words. The score does not have a lower bound. For reference, an undergraduate should receive a score between 30 and 50 (Kincaid et al. 1975). The Flesch-Kincaid Grade level is made popular by standardized school testing, library book classification, and its implementation in Microsoft Word. The score is computed by

$$0.39\left(\frac{totalWords}{totalWords}\right)+11.9\left(\frac{totalSyllables}{totalWords}\right)-15.49$$

(Kincaid, Fishburne, Rogers, & Chissom, 1975, p. 14).

McLaughlin (1969) created the SMOG readability score to address "the degree to which a given class of people find certain reading matter compelling and comprehensible" (DuBay, 2004, p. 3). To calculate the score, he simply counts the number of syllables in each word in a sentence and plugs them into the formula $SMOG = 3 + \sqrt{\#\, polysyllableWords}$ (DuBay, 2004, p. 47). Other less popular scoring methods include the Coleman-Liau index (CLI) and the ARI, which rely on characters instead of syllables per word. The CLI is computed through

$$CLI = 0.0588L - 0.296S - 15.8$$

where

$$S = \frac{totalSentences}{totalWords} \times 100 \quad \text{and} \quad L = \frac{totalLetters}{totalWords} \times 100$$

(Reck & Reck, 2007, p. 6),

and the ARI is found by

$$ARI = 4.71\left(\frac{characters}{word}\right) + 0.5\left(\frac{words}{sentence}\right) - 21.43$$

(Reck & Reck, 2007, p. 5).

According to DuBay (2004), some educational researchers tend to discredit a single formula to evaluate student performance without consulting other grading methods and writing styles; however, the improved results achieved when incorporating them warrant their use. The major point of dissension is that the scores are inconsistent when assigning the appropriate level grade (DuBay, 2004, p. 55); however, although the numerical classifications do have some variation, the difference in assignments do not separate significantly beyond educational categories (e.g., middle school, high school,

college). Even with all the opposition of readability statistics they are the most "accurate grade level predictors and difficulty classifiers" (DuBay, 2004, p. 34) and significant contributor to evaluation prediction (Ghose & Ipeirotis, 2011).

For the purposes of the classification, we use a normalized mean of the different readability scores, White's research uncovers "easy-reading text improves comprehension, retention, and reading speed, and that the average reading level of the U.S. adult population is at the eighth grade level" (White, 2003). Accounting for variability, we classify a score between 7th and 9th grade being optimal, less complicated being worse, and more complicated being the worst.

2. General Statistics

Although correlated with the readability statistics, we take word counts and character counts as an additional predictive variable to determine whether RSs or ROs expand on the quality of MROs performance. Zive (2015) proposes that using word and character length provides insight into the complexity of the documents.

3. Spell Check

Prior to processing the data to create the corpus, we check the spelling of the fitness report comments to ensure each word is counted appropriately. Ghose and Ipeirotis indicate through their research that spelling errors in reviews are a strong indicator of positive and negative tones in product reviews (Ghose & Ipeirotis, 2011). While A-PES offers the options of a spell check, each fitness report still contains spelling mistakes and can be used either to determine whether there is positive or negative sentiment in the comments, or the amount of effort a reporting senior is willing to commit to a MRO. Rinker developed an **R** package **qdap "**to bridge qualitative transcripts of dialogue and statistical analysis and visualization" (Rinker, 2013). We use **qdap** for two purposes: 1) actually conduct the spell check and replace the incorrect words, and 2) count the number of misspelled words in each fitness report. The military's propensity of using acronyms and professional jargon reduces the fidelity on the appropriate selection for an automatic spell check. As a result, we use a human-in-the-loop approach to develop a dictionary of suitable correction for non-standard words.

4. Non-Parametric Statistics

We examine now whether spelling errors, word counts, character counts, and the five readability statistics are independent of the fitness report's corresponding mathematical tier. For example, if spelling errors decrease between MRO tiers, we can test the hypotheses about means, of the form null hypothesis:

$$H_0 :\ \theta_{topThird} = \theta_{middleThird} = \theta_{bottomThird}$$

and alternatively, hypothesize:

$$H_1 :\ \theta_{topThird} \leq \theta_{middleThird} \leq \theta_{bottomThird} \text{ or } \theta_{topThird} \geq \theta_{middleThird} \geq \theta_{bottomThird}$$

where at least one of the inequalities is strict.

For this we use the Jonckheere and Terpstra (J-T) test (Sprent & Smeeton, 2007), which is a non-parametric one way analysis of variance rank test for populations that are ordered. It is an adaptation of the Kendall's τ rank correlation test and uses a normal approximation of the form

$$z = \frac{(U - E(U))}{\sqrt{Var(U)}},$$

where U is the rank correlation. Approximate normality is valid provided that the sample sizes that form each population are sufficiently large. To calculate the J-T test statistic we use the R function jonckheere.test() in the **clinfun** package (Seshan, 2016). When examining the results of this or any statistical test, we are mindful that with large sample sizes such as in the present study, even small effect sizes can produce p-values that are extremely small. Rejection of the null hypothesis, therefore, may not indicate practical importance.

5. Word Correlation Analysis

The analysis of word association uses the Pearson's correlation coefficient which is calculated by

$$\rho_{X,Y} = \frac{E\left[XY\right] - E[\mathrm{X}]\mathrm{E}[\mathrm{Y}]}{\sqrt{E[\mathrm{X}^2] - [E[X]]^2}\sqrt{E[Y^2] - [E[Y]]^2}}$$

Suggested by Evans (1996), we classify a term as having using the following scales: 0.00-0.19= "very weak," 0.20-0.39= "weak," 0.40-0.59= "moderate," 0.60-0.79= "strong," and 0.80-1.0= "very strong" (Evans J. , 1996). Although correlation can be positive or negative, in this study we focus on those terms that have positive correlations with specific words.

D. MODELING

The modeling phase of our study consists of two components: predictive modeling and natural language processing modeling of word association. As outlined in Figure 6, our process takes each term-document matrix through a series of selection processes to generate an optimized model.

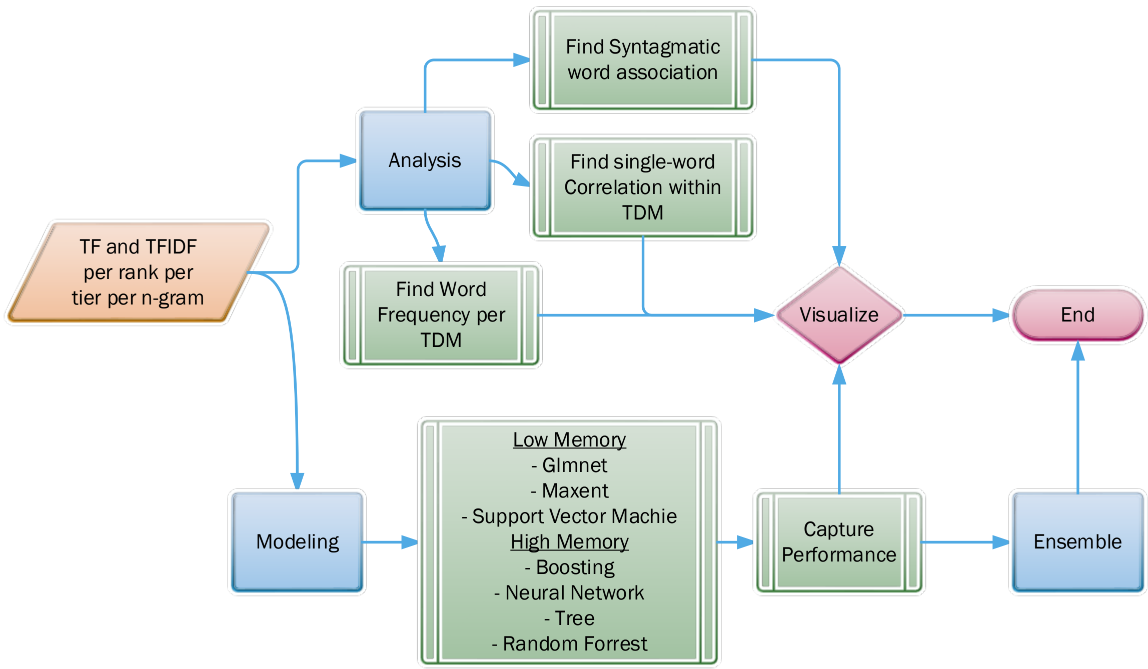


Figure 6. Modeling Process Map

During predictive modeling, we take each of the 360 document-term matrices and run them through seven supervised learning models: 1. Lasso and Elastic-Net Regularized Generalized Linear Models, 2. Classification and Regression Trees, 3. Support Vector Machines, 4. Boosting, 5. Random Forests, 6. Neural Networks, and Maximum Entropy. These matrices represent the predictive variables and the response variable is the tier associated to the relative value for RSs or comparative assessment for ROs. We capture the best document-term matrix configuration and run them through a stacking ensemble using a generalized linear model.

1. Lasso and Ridge Regularized Generalized Linear Models

Using spelling errors, word count, character count, ARI score, Coleman-Liau score, SMOG index, Flesch - Kincade Index, and Flesch grade level we attempt to predict the performance tier through a variety of models. We start with the Lasso and Ridge Regularized Generalized Linear Models (elastic net) which takes a target vector

$\mathbf{y} = (y_1, \ldots, y_j)^t$ and outputs

$$\hat{\mathbf{y}} = (\hat{y}_1, ..., \hat{y}_j)^t \quad \text{where } 0 \le \hat{y}_j \le 1 \text{ and } \sum \hat{y}_j = 1.$$

The multinomial logistics regression uses

$$\hat{\eta}_j = \hat{\beta}_{0j} + \hat{\beta}_{1j} X_1 + ... + \hat{\beta}_{kj} X_k$$

$$\hat{y}_j = \mathrm{e}^{\hat{\eta}_j} / \sum_{l=1}^{k} \mathrm{e}^{\hat{\eta}_l}$$

with a regularized score function of

$$\max_{\beta_0, \beta} \left\{ \sum_{i=1}^{n} \left[y_i (\beta_o + \beta^T x_i) - \ln(1 + e^{\beta_0 + \beta^T x_i}) \right] - \lambda \sum_{j=1} \left| \beta_j \right| \right\}$$

(Hastie, Tibshirani, & Friedman, 2009, p. 125).

2. Classification and Regression Trees

We next use classification and regression trees to predict the tier ranking. This model is useful because it does not tend to over fit the model by including irrelevant predictors and it handles missing values well (Feund & Shapire, 1995). For a categorical variable with *J* levels, the target is a vector ***y*** with *J* binary variables $\mathbf{y} = \left(y_1, y_2, ..., y_J \right)^t$. If the i^{th} observation falls in the m^{th} leaf then the output gives the distribution of the y's in the m^{th} leaf

$$\hat{\mathbf{y}} = \left(\hat{y}_1(\mathbf{x}_i), \hat{y}_2(\mathbf{x}_i), ..., \hat{y}_J(\mathbf{x}_i) \right) = \left(\frac{n_{1m}}{n_m}, \frac{n_{2m}}{n_m}, ..., \frac{n_{2m}}{n_m} \right) \quad \text{for } \mathbf{x}_i \in R_m.$$

For multinomial classification, the score function in **rpart** is the weighted sum of the leaf impurities

$$Score = -2\sum_{m=1}^{M} n_m \sum_{j=1}^{J} \frac{n_{jm}}{n_m} \log\left(\frac{n_{jm}}{n_m} \right) = \sum_{m=1}^{M} n_m \left[-2\sum_{j=1}^{J} \hat{y}_{jm} \log\left(\hat{y}_{jm} \right) \right]$$

(Therneau, Atkinson, & Ripley, 2015).

3. Support Vector Machine

The idea of support vector machines is to re-imagine classification based on extending linear decision boundaries to non-separable classes. The SVM becomes a non-linear optimization problem that for each $i \in \{1, ..., n\}$ we introduce a variable $\zeta_i = \max(0, 1 - y_i(w \times x_i + b))$ where ζ_i is the smallest non-negative number satisfying $y_i(w \times x_i + b) \geq 1 - \zeta_i$. The resulting primal formulation is

$$\min \left[(1/n) \sum_{i=1}^{n} \zeta_i \right] + \lambda \|w\|^2$$

subject to $y_i(w \times x_i + b) \geq 1 - \zeta_i$ and $\zeta_i \geq 0 \forall i$. The problem is solved by taking the Lagrangian dual and maximizing

$$f(c_1, ..., c_n) = \sum_{i=1}^{n} c_i - 1/2 \sum_{i=1}^{n} \sum_{j=1}^{n} y_i c_i \left(x_i \times x_j \right) y_j c_j$$

subject to $\sum_{i=1}^{n} c_i y_i = 0$ and $0 \leq c_i \leq 1/(2n\lambda) \forall i$ where c_i is defined such that $w_i = \sum_{i=1}^{n} c_i y_i \vec{x}_i$ (Burges, 1998).

4. Boosting

Boosting is an extension of the tree model developed by Freund and Shapire (1997). It uses an algorithm called AdaBoost, which combines the weighted sum of repeated weak learner models. A weak learner is one that “performs just slightly better than random guessing” (Feund & Shapire, 1995). The algorithm maintains a distribution of weights over the training set and calculates the goodness of a weak model through its error

$$\varepsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i] = \sum_{i: h_t(x_i) \neq y_i} D_t(i)$$

where t is the round in set T, y_i is the label in set Y, x_i is a feature in the domain space X, and finally $D_t(i)$ is the weight of the distribution on example i on round t. In the first round, all weights are naively set to be the same. In subsequent rounds, the "weights of the incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set" (Freund & Schapire, 1999, p. 772).

5. Random Forest

Random Forest is another variant of the tree model developed by Breiman (2001). The model falls under the ensemble idea where flexible models generally have low bias and high variance. By averaging a lot of models, we can reduce the variance and the bias (Breiman, 2001). The idea is that we take a bootstrap sample and run a tree. At each split, we select a new random sample and select the best split. After running a bunch of pruned trees, a few strong predictors might be chosen on every tree and therefore would be assigned a higher variable importance. The model trains on a subset of the data and tests for accuracy on its complement (Breiman, 2001).

6. Neural Networks

The idea of a neural network is to "extract linear combinations of input features (the predictor variables) and then model the target as a nonlinear function of these" (Whitaker, 2017e). For a set of input nodes x_j in **X** with a bias node x_o, we take k+1 weights $v_o, v_1, v_2, \ldots, v_k$ and reduce the dimensionality of the problem. Each input is reduced to a "neuron" in a hidden layer. The neurons are a linear combination of the input nodes **X,** where $z = \sum_{j\geq 0} v_j x_j$. Now the bias node's weight acts as an activation function to move towards the output node where

$$\hat{y} = \begin{cases} 1 & \text{if } \sum_{j\geq 1} v_j x_j > -v_0 \\ 0 & \text{if } \sum_{j\geq 1} v_j x_j \leq -v_0 \end{cases}$$

for a specific classification prediction (McCullough & Pitts, 1943).

7. Maximum Entropy

The maximum entropy classification technique is a probabilistic model that does not assume that the features are conditionally independent of each other. Based on the principle of maximum entropy, the model selects from all the training data, the model that has the largest entropy (Vryniotis, 2013). Starting with the bag-of-words approach, each document in the corpus is represented with an array of 1s and 0s conditioned on whether a word w_i exists in the context of the document (Pang, Lee, & Vaithyanathan, 2002). The objective is to construct a stochastic model with contextual information as input x and an output value of a class y (Berger, 1996). We do this by selecting a distribution p from a set of allowed distributions that $\arg\max_{p \in P} H(Y \mid X)$ where

$$H(Y \mid X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y \mid x) \text{ (Wellner, 2017).}$$

Because entropy is measure of uncertainty, we can use it as a quality measure: the "more we know about something the lower the entropy" (Wellner, 2017); therefore, the more our model captures the structure of the language, the lower entropy.

8. Ensemble Modeling

Flexible "strong learner" models may have low bias, but are highly susceptible to variance (Whitaker, 2017b). Instead of relying on a single model, we can build an ensemble. An ensemble refers to "combining large sets of large classifiers built with randomness applied to data or classifier" (Buttrey, 2016b). If the fitted models are independent, we can consider building an ensemble of weak learners that maintain low variance and then averaging them, which reduces the variance of the prediction and results in a smaller prediction error (Hastie, Tibshirani, & Friedman, 2009, p. 605). The idea is analogous to voting where many weak voters who can predict the right candidate 51% of the time are better than a single strong voter who predicts 90% of the time.

There are multiple ways to using ensemble modeling. Boosting and bagging are mentioned above, but there is an additional way called stacking (Brownlee, 2016). The idea is to find models that are skillful in different ways, and allow a new classifier to find

the best prediction from each model and improve classification. In this case, the an additional classifier is a generalized linear model.

Although unrealistic for a promotion board member to mentally utilize an ensemble when developing a sentiment on the textual context of fitness reports, it offers insight on the potential predictability of fitness reports.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. MODEL RESULTS

This chapter discusses the results of our analysis. Our motivation is to be able to develop and use a variety of techniques to predict the performance tier.

We start by analyzing the quality of the response variable. Although MROs are assigned a tier, the concentration of marks in a few scores make it more difficult to separate the bottom and the top from the middle. We compare the intended distribution with the actual distribution. We further examine the response variable by determining how often the RO and RS assigned score place the MRO in the same tier given that the RO concurs with the RS markings.

We then analyze the text-statistics such as word count, character count, readability indices, and spelling errors to find patterns that enhance the informational value of the words. We use a penalty enhanced generalized linear model and classification regression trees with each of these predictive variables to train a model and test it against the validation set. Our objective is to find the elements that are practically significant for prediction.

Examining the words, we investigate compliance with the PES manual by highlighting when the RS writes a mandatory directed comments. We select five easily subset-able comments and run regular expressions queries to find the proportion of reporting seniors that comment on the appropriate directed comment.

When spell checking, we discover that many RSs and ROs like to qualify the MRO's performance by making a ranking based statement such as "#1 [rank]…" or "Best [rank]..." We consider instances when the statement is used and determine whether the MRO is the top ranked Marine at the time of processing and cumulatively.

The next part examines which terms are correlated with "promotion," "potential," "retention," "assignment," "command," and professional military schools. We use three separate techniques to find these words: supervised correlation mapping, unsupervised correlation mapping, and a neural network.

Finally, we run a series of predictive models to find which document-term matrix configuration has the most predictive power. By isolating the best ones, we can pull the useful predictive variables and their magnitude. These variables represent the most powerful words in predicting the MRO's performance tier based on the textual information. We pull the five best text-based predictive models and combine them with the two text-statistics based models to form an ensemble.

A. PRE-CORPUS ANALYSIS

1. Analysis of the Relative Value and Comparative Assessment

To analyze the quality of our response variables we analyze their distributions and whether to RS and RO quantify the MRO in the same performance tier group.

a. Analysis of the Relative Value and Comparative Assessment Distribution

Because the performance tier group is our response variable, we investigate the quality of variable by comparing the intended distribution with its actual one. The Basic School does not explain how to generate relative values and gives an incorrect impression that the RV automatically normalizes fitness report averages into a "bell curve" distribution (Clemens et al., 2012). The future RSs are taught that the lower bound is 80 and the upper bound is 100 (Dodd, 2016). The perceived bell curve would best be exemplified by a censored normal distribution with displayed in Figure 7.

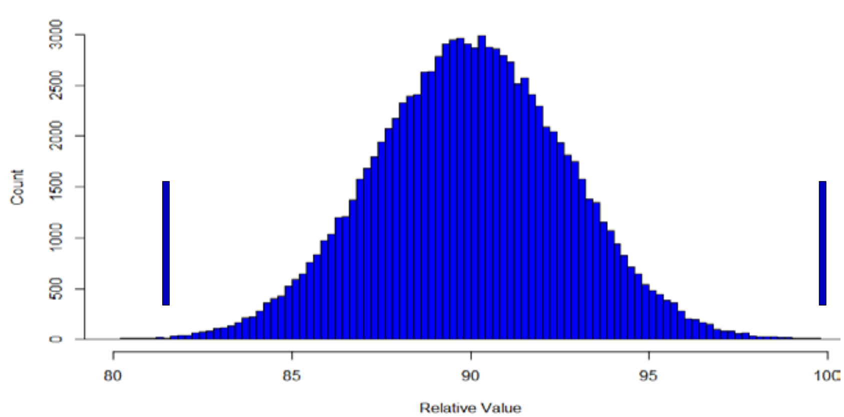


Figure 7. Sample Censored Histogram of Relative Value

By plotting the sample distribution of the relative value scores in Figure 8, we see that the distribution reflects the maximum property with high occurrences at 90 and 100, and appears to be more triangular in between with a min of 80, mode of 90, and max of 100, which reflects the true relative value distribution equation of

$$\max\left(80, \frac{RawScore - RSaverage}{RS\max - RSAverage} * 10 + 90\right),$$

where the raw score is the individual FitReps's average, the RS average is the average of all the RS's raw scores, and the RS max is the highest raw score an RS assigned to an MRO.

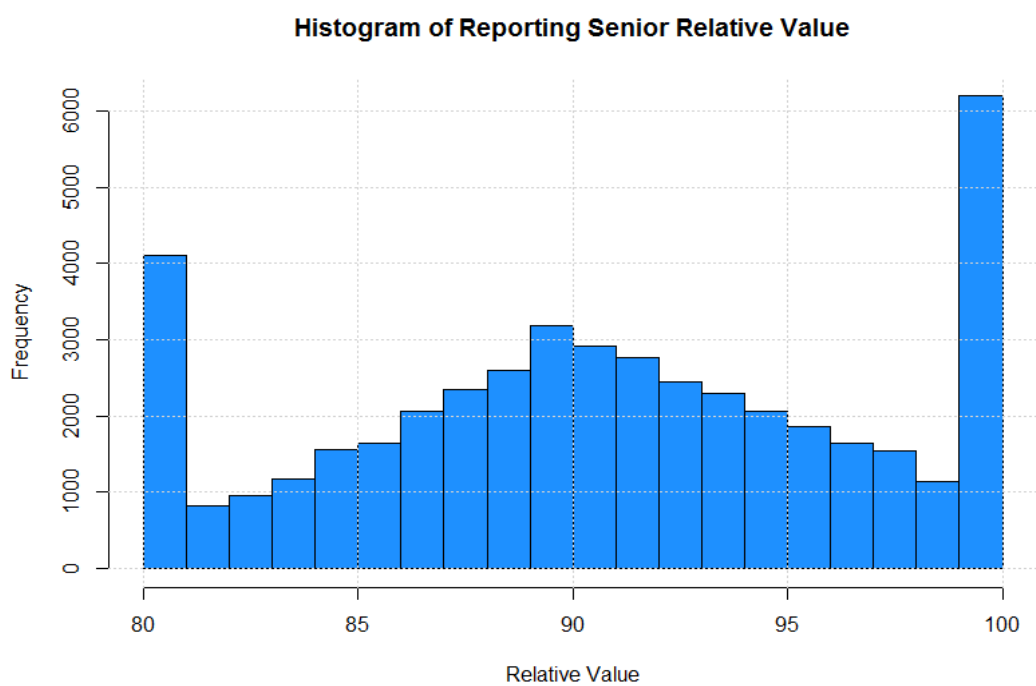


Figure 8. Histogram of Relative Value Scores

We notice that there appear to be more observations greater than 90 than less. In our sample size of 50,267 fitness reports, 30,198 were greater than 90 while only 20,069 were less. We also notice that there are more 100s than 80s. When we count those values

we find that there are 5,776 100s for 3,568 80s. We conclude that the data have a higher concentration in the upper half than the lower half.

This distribution however does not tell the full picture. We examine the mean, mode, percent above the mean, percent with or below the mean, percent at 80, and percent at 100. Our results shown in Table 3 demonstrates that relative values tend to increase as the rank of the MRO increase.

Table 3. Relative Value Summary Statistics by Rank

Rank	Mean	Median	% ≤ 90	% > 90	% at 80	% at 100
2ndLt	90.31	89.73	58.06	41.94	7.43	13.24
1stLt	90.59	90.45	52.46	47.54	7.39	11.21
Capt	90.93	90.90	44.63	55.37	6.37	10.83
Maj	91.23	91.42	38.78	61.22	8.44	11.6
LtCol	91.81	93.04	34.14	65.86	5.11	13.19
Aggregate	90.81	90.78	39.99	60.01	7.10	11.49

To visualize our results, we plot the estimated cumulative distribution function of each rank. Figure 9 demonstrates that relative values tend to increase as the rank of the MRO increase. To determine whether this could happen by chance alone, we use the Jonckheere and Terpstra test with an alternative hypothesis that assigned marks increase by ranks. The test overwhelmingly rejects the null-hypothesis with a p-value less than 2.2×10^{-16}. This shows that the relative value tend to concentrate on one side or the other.

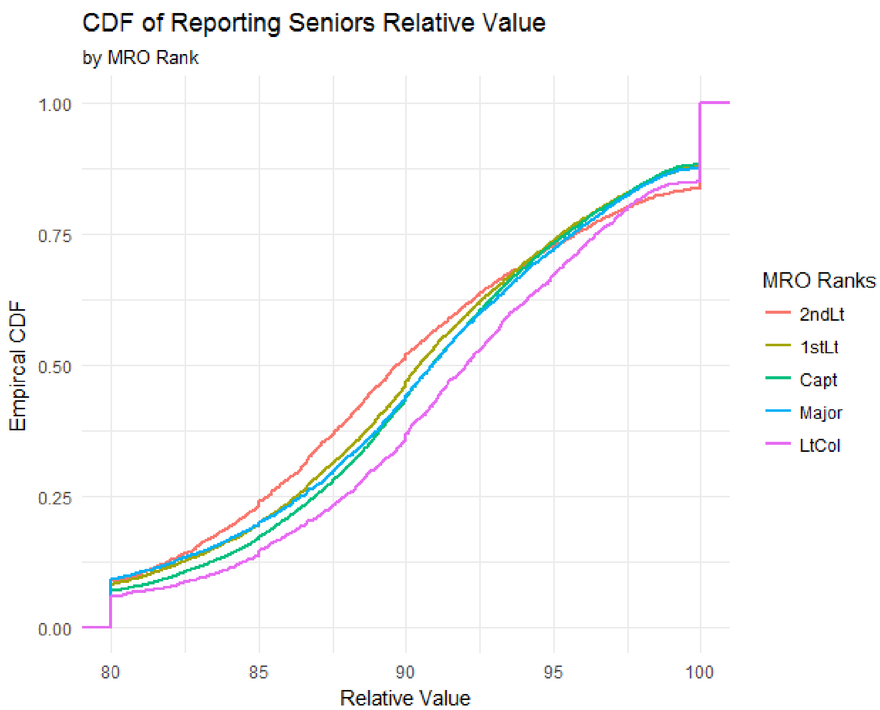


Figure 9. Estimated Cumulative Distribution Functions of Reporting Senior Assessments

The trending profiles demonstrate concentrations that depart from the expected distribution. These concentrations tend to lose information value on the quality spread and affect the predictability of word pictures.

The skewness in the relative value is potentially due to human error when reviewing the MROs. As MROs attrite out of the system, the RS and RO generally observe officers to be closer to the Marine Corps values consistent with their promotions. At the lower and upper ranks, the concentration of MROs in a few evaluation points reduce the Marine Corps's ability to separate talent or ineptitude from the mass because everyone is quantitatively characterized equivalent. Furthermore, when analyzing the value gained from text, the narrower spread makes predictability less reliable as relative values overlap between tiers.

We also examine the RO's comparative assessment distribution to determine how reliable our Section K response variables are. The comparative distribution trends by rank are the same as the relative value with a further observation: there is a disparity between

the guidance and practice. Contrary to the relative value which is automatically calculated based on the raw score, high mark, and average, the comparative assessment is an unconstrained fixed mark. Taken from an actual FitRep, Figure 10 shows the guidance for assigning the relative comparison marks to the MRO. The marks are supposed to resemble a "Christmas Tree" with few occurrences on the top and the very bottom, but a gradual accumulation of observations as one descends the tree (Commandant of the Marine Corps, 2015).

Figure 10. Distribution Guidance for Reviewing Officer Comparative Assessment. Source: The Basic School (2016c, p. 14).

Clemens et al. analyze the distribution of RO marks over time and conclude that the marks do not match the intended distribution and suggest that "they are less informative about the true spread of quality than they are intended to be" (Clemens et al., 2012, p. 14). Figure 11 displays their conclusion and shows that lieutenant colonels get ranked higher than second lieutenants.

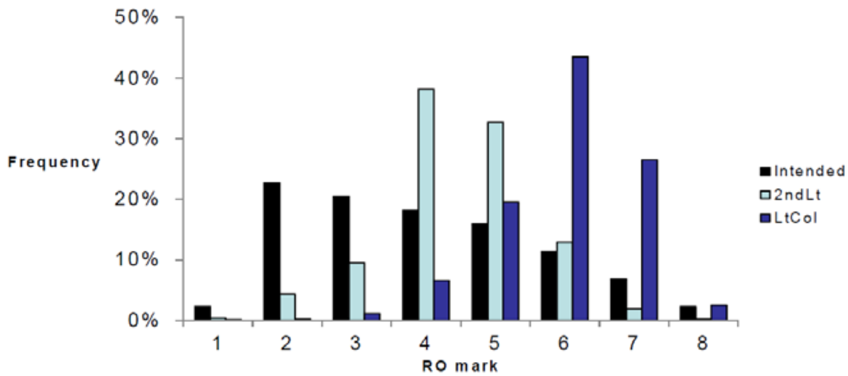


Figure 11. Distribution of RO marks among second lieutenant and lieutenant colonel FitReps compared with the intended RO mark distribution within a grade. Source: Clemens et al. (2012, p.15).

To further investigate this discovery, we plot the cumulative distribution function of each rank compared against the intended "Christmas Tree" distribution and against the average. Figure 12 clearly demonstrates a departure from the intended distribution and a gradual increase by rank from second lieutenants to lieutenant colonels with captains being closest to the mean; however, all the estimates are subject to standard errors.

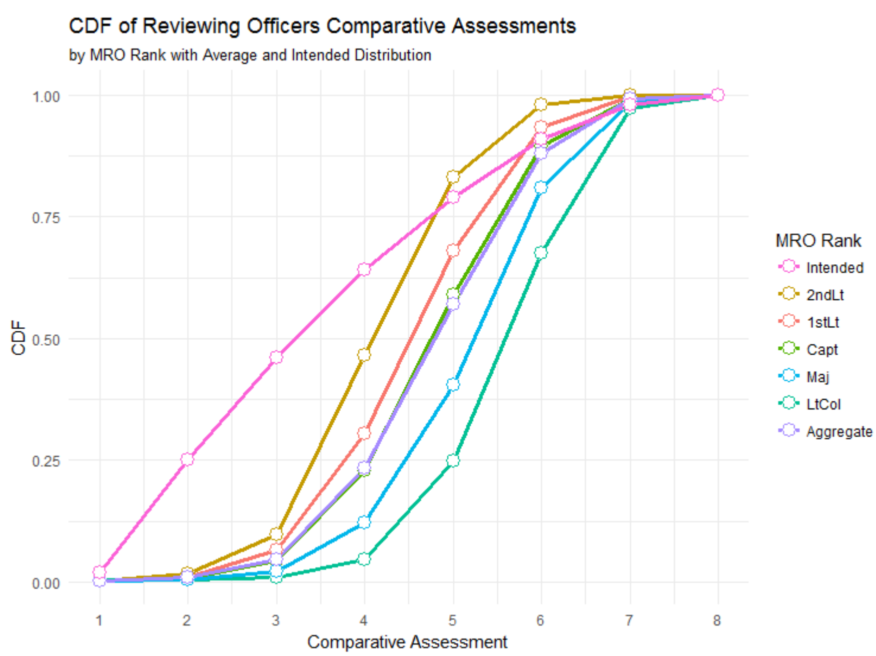


Figure 12. Estimated Cumulative Distribution Function of Comparative Assessment

We use the Jonckheere Terpstra test as with the relative value and reject the null-hypothesis that

$$CAMean_{2ndLt} = CAMean_{1stLt} = CAMean_{Capt} = CAMean_{Maj} = CAMean_{LtCol}$$

with a p-value of 2.2×10^{-16}. These are comparative assessments and the marks are projected to the promotion board members as those above, with, and below the MRO. Assigning a higher score does not help the individual and, coupled with the departure from the intended distribution, the RO's marks loses informational value on the MRO's performance.

This departure from the intended distribution makes quantifying the RO's assessment of the MRO difficult to interpret. For example, almost half the lieutenant colonels are marked in block 6 and 73% are marked in blocks 6 and 7. This small dispersion between blocks makes it not only difficult to quantify the quality of the MRO's performance, but matching appropriate comments to a response variable near impossible. The comparative assessment is meant to be a comparison between all officers in the same grade as the MRO. The gradual bunching of officers towards the top of the grading scale makes everyone average when examining the scores in an aggregate.

We conclude that while we are not affected by the increasing trend by rank of the response variable, we are affected by the lack of dispersion and presence of concentration in the relative value and comparative assessments. This concentration makes a clear separation between thirds difficult to model and leads to a general misclassification of fitness reports towards the middle tier.

b. Analysis of Concurrence Between RS and RO Assessments

We further investigate the quality of our response variable by analyzing how often the RS's and RO's marks place the MRO in the same tier. Section K, Item 2 requires the reviewing officer to declare whether he concurs with the reporting senior's numeric and qualitative evaluation (Commandant of the Marine Corps, 2015). The RO must also closely scrutinize all reports and assess the RS's execution of his reporting

responsibilities. Ideally, the RO and RS would assign the MRO to the same block or, for the misaligned tier blocks, we would observe a proportional amount of "non-concur."

In the FitRep sample of 71,212, only 105 or less than 0.1% did not concur with the evaluation; however, as demonstrated by Table 4, there is a 49% disparity between assigned tiers. We calculated the disparity by taking

$$1 - \frac{TrueNegative + TruePositive}{Total},$$

which simply means, what proportion of time the tier groups are not the same.

Table 4. Reviewing Officer Versus Reporting Officer Tier Assignments

		Reviewing Officer			
		Tier 1	Tier 2	Tier 3	Total
Reporting Senior	Tier 1	8,868	4,920	1,595	15,383
	Tier 2	4,877	7,008	5,228	17,113
	Tier 3	1,040	3,662	6,139	10,841
	Total	14,785	15,590	12,962	43,337

By taking the absolute value of the differences, the data show that they disagree by one tier assignment 43.1% of the time, and by two tier assignments 6.1% of the time. We examine the differences in Table 5 and find that the ROs mark the MRO higher than the RSs 27.1% of the time and the RSs higher than ROs 22.1% of the time.

Table 5. Tier Assignment Differences Between RO and RS

RO Evaluates Higher		Concur	RS Evaluates Higher	
-2	-1	0	1	2
0.037	0.234	0.508	0.197	0.024

Some blending of the borders is expected between thirds; however, with 99.9% of ROs concurring with the RSs' observations, MROs should expect better harmony between the

RS and RO's marks. Through the central limit theorem, we know that the large sample size displays proportions that are very close to the truth.

One of the primary purposes of the concurrence requirement was to prevent inflation and address adversity (Commandant of the Marine Corps, 2015); however, there is little guidance on how to address disparity within the evaluation chain. Of note, there is no requirement to mark a fitness report adverse if there is non-concurrence; the RO simply has to comment on why there is discrepancy.

The discrepancy can be explained for a variety of reasons. The RS marks the report based on the MRO's performance during the reporting period. The RO marks the report based not only on performance, but also on comparing the MRO to all Marines of that grade known professionally to the RO (Marucci, 2016). While the MRO's relative value gets recalculated as the RS writes more fitness reports, the RO's assessment is fixed. Some reporting seniors and reviewing officers may not understand how the system works. Low number of initial reports for the RS or RO may force an MRO into an undeserving tier.

2. Descriptive Statistics of Corpus

In this section we pursue the research by Ghose and Ipeirotis (2011) and Jordan (2011) on the information value of readability statistics, spelling errors, word counts, and character counts for the reporting senior and the reviewing officer. Because our end state is to predict performance tier groups based on information about or contained in the comments, we focus our analysis on the differences between these tier groups.

a. Word count and Character Count Analysis

We analyze whether there is informational value in the word and character counts by performance tier. Table 6 presents summary statistics on word and character counts and we observe that there is an increasing trend between Tiers between relative value-based tier classification are statistically significant.

Table 6. Word and Character Count Summary Statistics for Sections I and K

	Word Count			Character Count		
	mean	median	mode	mean	median	mode
RS						
Tier 1	140.70	146	148	766.76	804	855
Tier 2	135.51	141	149	743.13	782	850
Tier 3	126.88	133	145	694.44	734	845
RO						
Tier 1	78.78	83	91	426.20	457	496
Tier 2	72.92	77	87	400.36	426	493
Tier 3	68.93	72	84	380.12	400	495

Figures 13 through 16 confirm this apparent trend between the word and character counts in sections I and K comments.

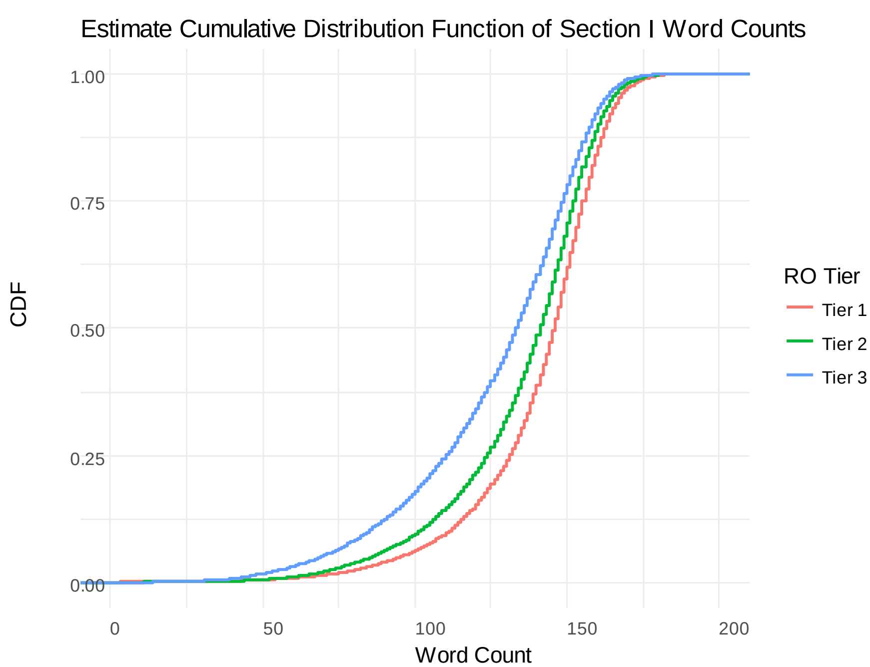


Figure 13. Cumulative Distribution Function of Section I Word Counts

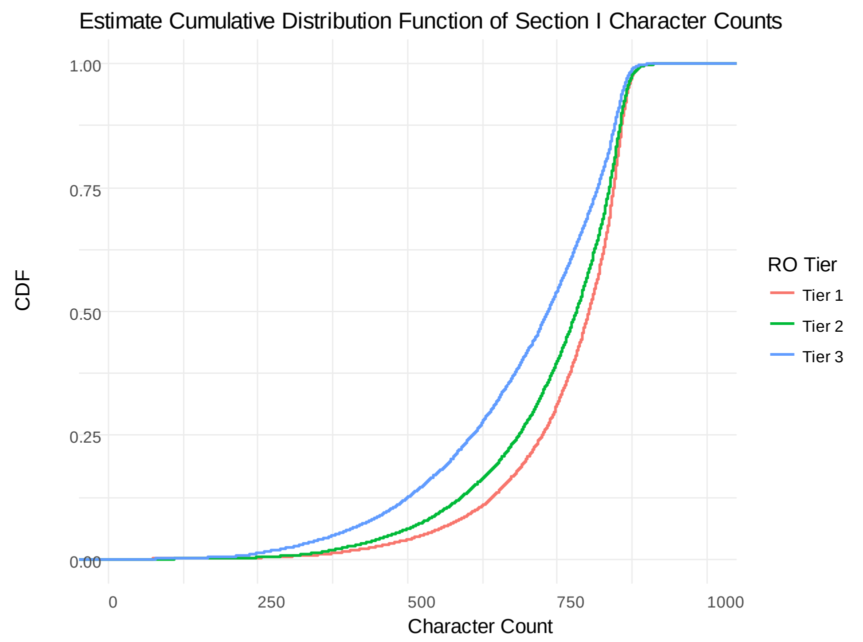


Figure 14. Cumulative Distribution Function of Section I Character Counts

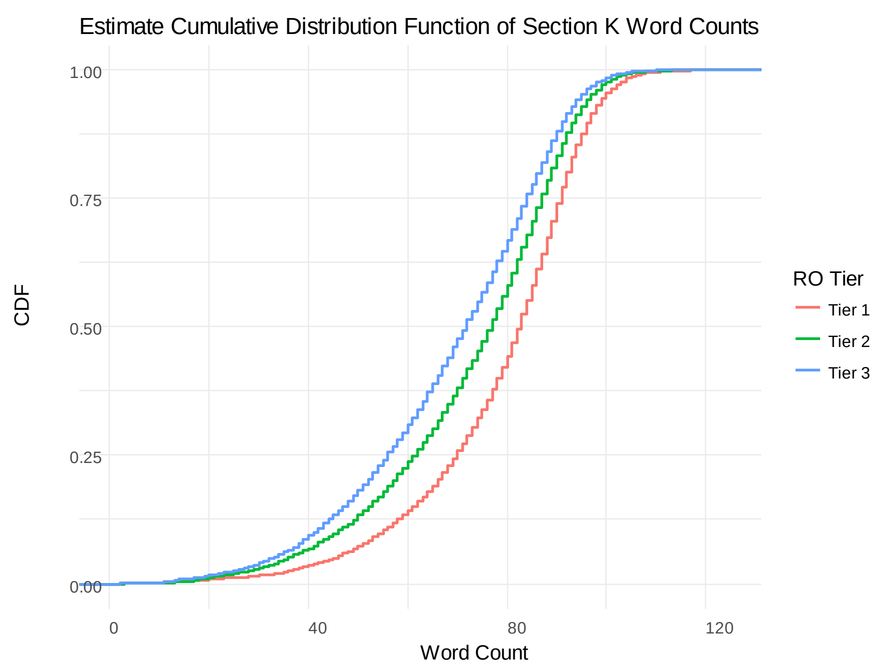


Figure 15. Estimate Cumulative Distribution function of Section K Word Counts

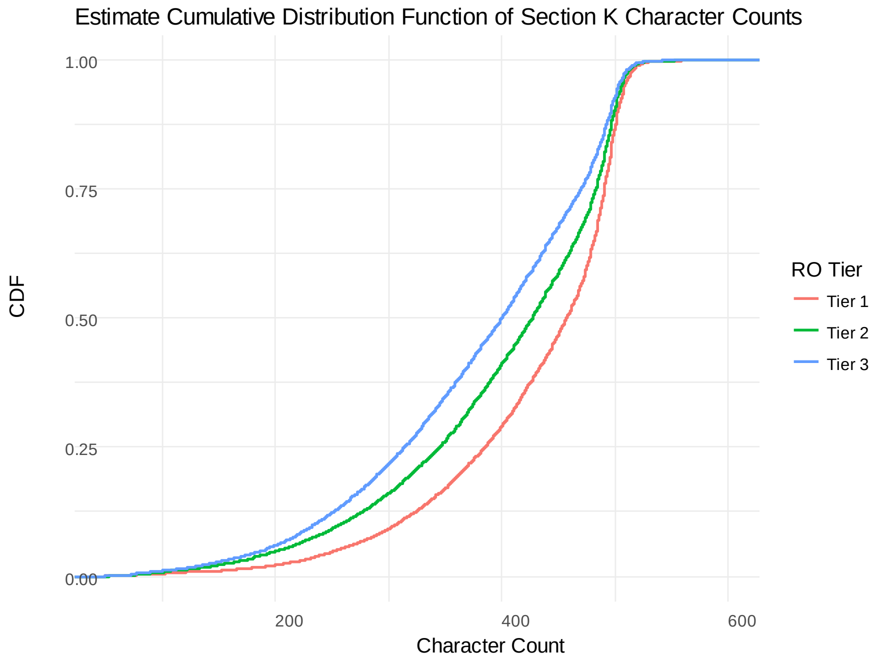


Figure 16. Estimate Cumulative Distribution function of Section K Character Counts

We use the Jonckheere and Terpstra test ordered by an increasing number of words per tier for sections I and K. Each test returned a p-value less than 2.2×10^{-16}. While the sample size would lead us to find statistical significance, we find this practically significant due to the consistent large difference between tiers.

b. Spelling Error Analysis

We find when analyzing the number of spelling errors that while there is statistical significance, the results are not practically significant. We start by analyzing descriptive statistics by tier of spelling mistakes by comment section in Table 7.

Table 7. Descriptive Statistics of Section I and K Spelling Errors

	Number Misspelled		
	mean	median	mode
RS			
Tier 1	0.469	0	0
Tier 2	0.427	0	0
Tier 3	0.343	0	0
RO			
Tier 1	0.46	0	0
Tier 2	0.39	0	0
Tier 3	0.36	0	0

The spelling errors are adjusted for acronyms and jargon during our 'man-in-the-loop' spell check process by developing an evolving dictionary. Interestingly, the APES user-interface provides the ability to use a spell checker automatically when routing is selected. The RS and RO have to decline its use prior to forwarding to the next element in the chain. We observe the estimated cumulative distribution functions to visualize our results in Figures 17 and 18.

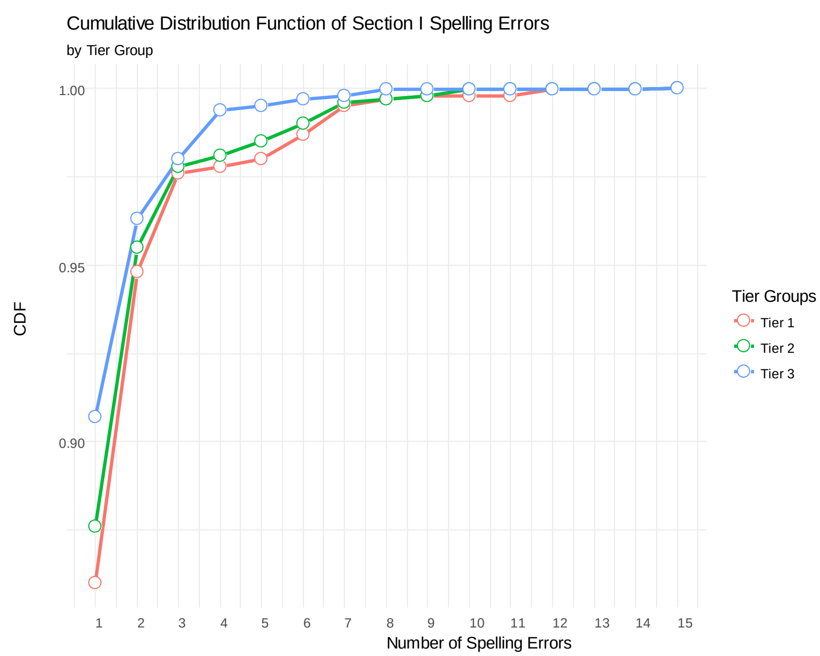


Figure 17. Estimate Cumulative Distribution Function of Section I Spelling Errors

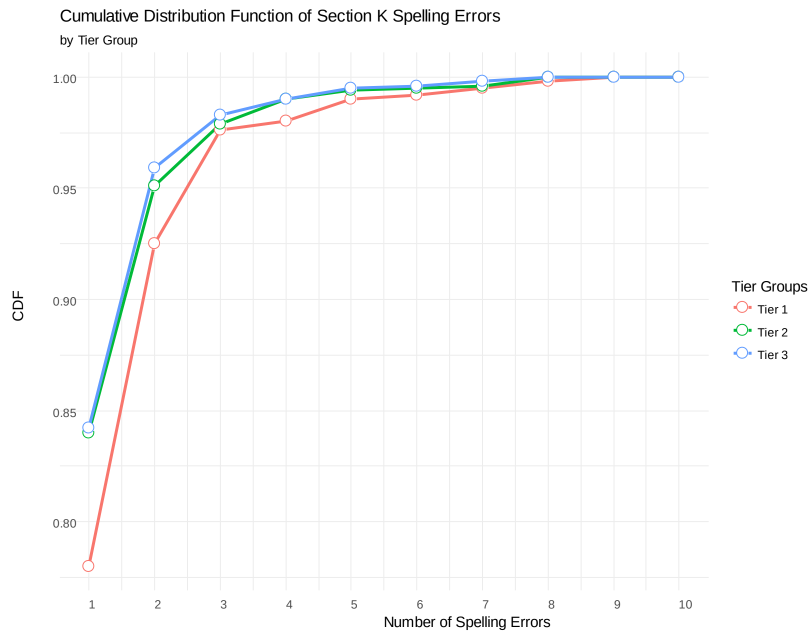


Figure 18. Estimate Cumulative Distribution Function of Section K Spelling Errors

The number of spelling errors exhibit statistical significance and differences between tiers with a Jonckheere and Terpstra test p-value less than 3.139×10^{-14} for the RS and 2.2×10^{-16} for the RO; however, this anomaly is principally due to longer comments afford the RS and RO more opportunities to make spelling errors. We find examination of the spelling errors do not provide additional information value on predicting tier assignments.

c. Readability Statistics

The readability statistic results were consistent with the sentiment analysis conclusions of Ghose and Ipeirotis (2011). Positive sentiment is generally associated with smaller, simpler writing. The models that are based on the number of syllables per word and words per sentence, such as SMOG, Flesch-Kincaid, and Flesch Grade level indicate that the comments become simpler and more readable as tiers go from three to one. Alternatively, models that are based on the number of characters per word such as the Coleman-Liau Index point toward an increase in complexity as FitReps go from tier three to tier one.

We start by analyzing the measures of central tendencies of each index. The results in Table 8 show very moderate increase in complexity between tiers one and three for the ARI, Flesch Kincaid Score, and SMOG Index. Conversely, the Coleman Liau Index increases in complexity between tiers three and one.

Table 8. Descriptive Statistics for Sections I and K Readability Indices

	ARI			Flesch Kincaid Score			SMOG Index			Coleman Liau Index		
	mean	median	mode	mean	median	mode	mean	median	mode	mean	median	mode
RS												
Tier 1	11.78	11.55	11.51	11.98	11.8	12.27	14.08	13.93	13.02	34.03	34	34
Tier 2	11.99	11.76	11.44	12.11	11.97	11.92	14.20	14.07	14.55	33.09	33	35
Tier 3	11.93	11.73	12.59	12.05	11.9	11.77	14.15	14.02	14.55	33.18	33	33
Aggregate	11.90	11.68	11.44	12.05	11.89	11.17	14.15	14.01	14.55	33.44	33	33
RO												
Tier 1	10.47	10.2	10.17	11.02	10.82	10.14	13.13	13.02	13.02	36.02	36	36
Tier 2	10.87	10.64	9.8	11.31	11.14	10.14	13.30	13.02	13.02	34.23	34	35
Tier 3	11.08	10.83	11.48	11.47	11.25	10.14	13.44	13.3	13.02	33.52	34	32
Aggregate	10.80	10.55	11.48	11.26	11.07	10.14	13.28	13.02	13.02	34.62	35	35

In analyzing Figures 19 through 28, we find that each one of these readability statistics show very little dispersion between tiers. Even when we shrink the limits to decipher thse important areas, the lines overlap.

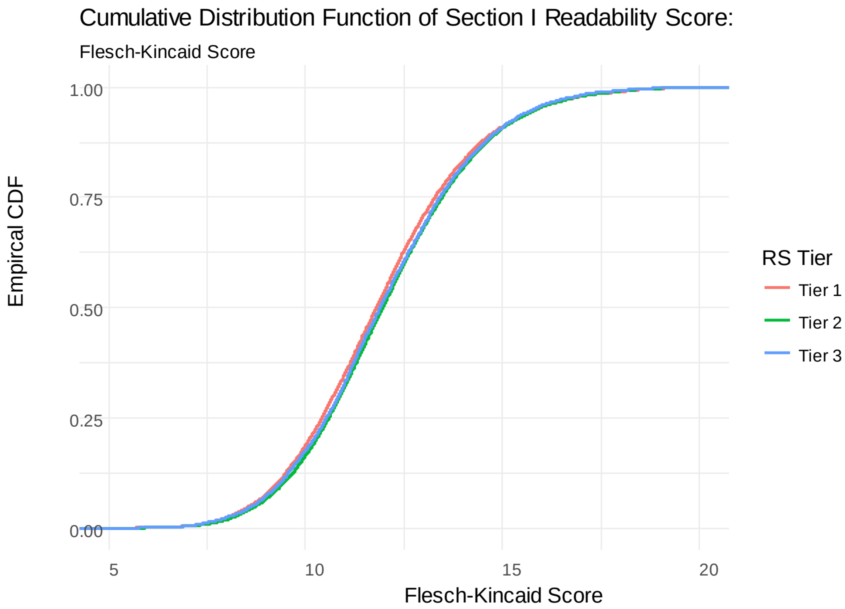


Figure 19. Cumulative Distribution of Section I Flesch-Kincaid Readability Score

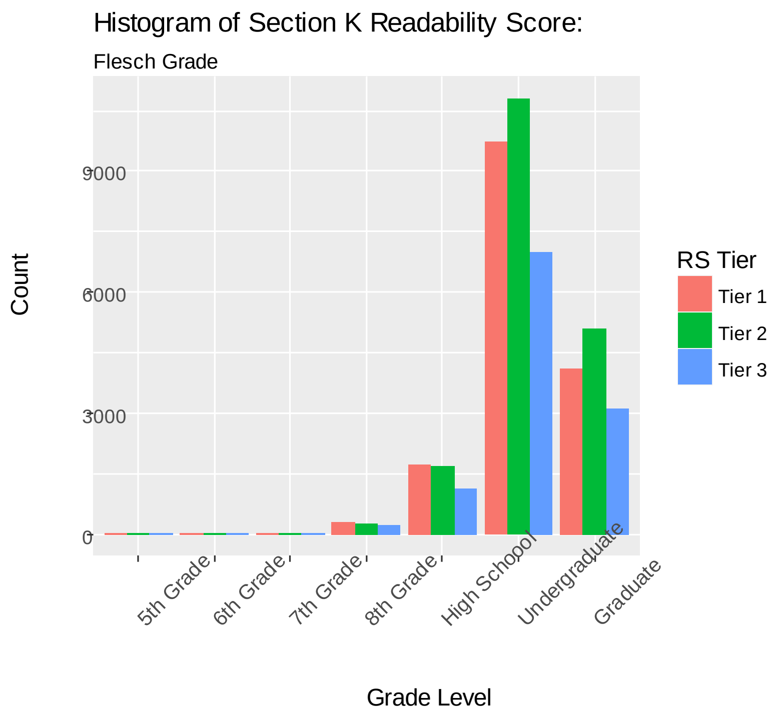


Figure 20. Histogram of Section I Flesch Grade Level

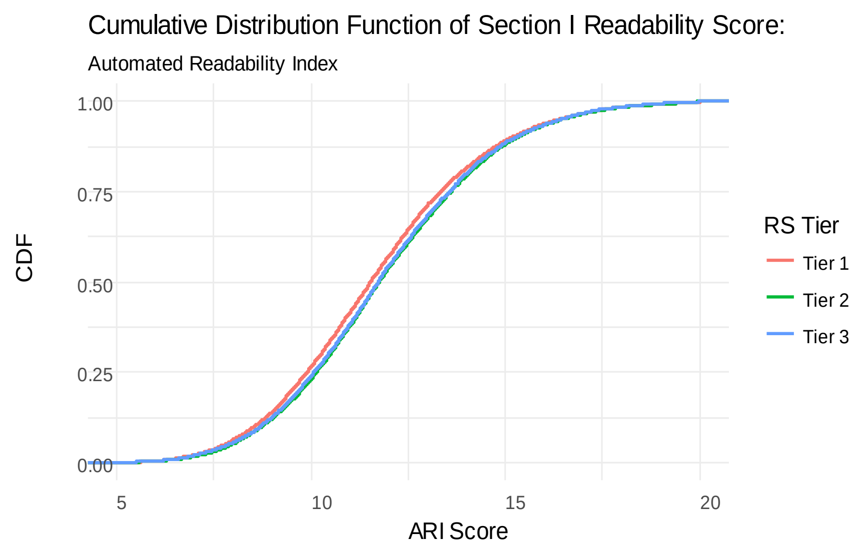


Figure 21. Cumulative Distribution of Section I ARI Readability Score

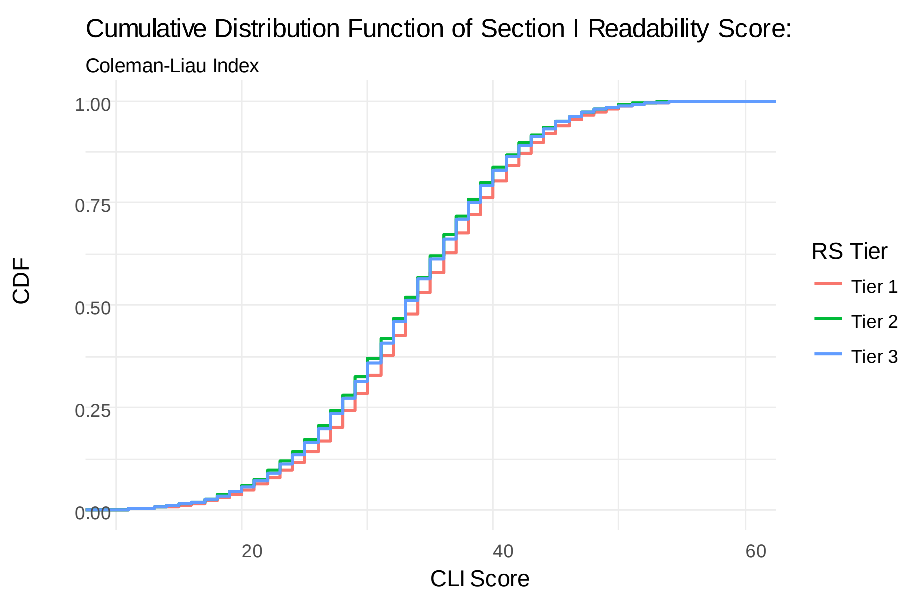


Figure 22. Cumulative Distribution of Section I Coleman Liau Readability Score

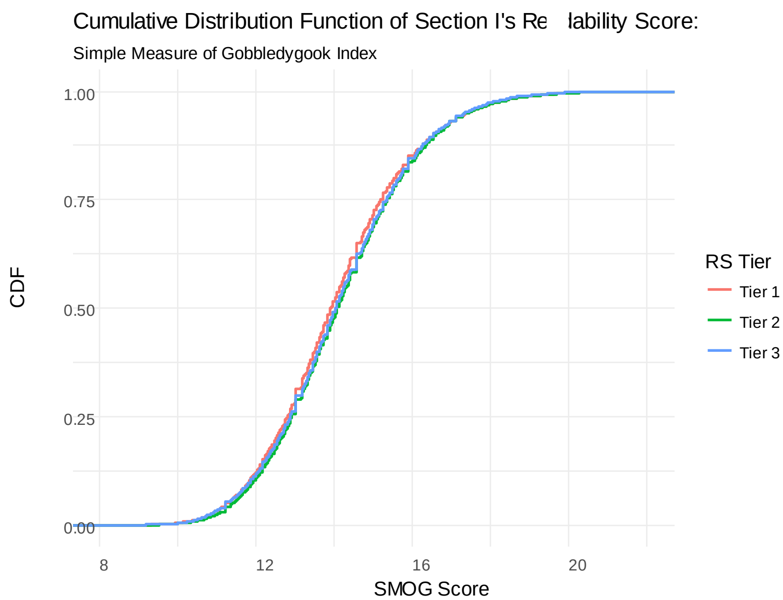


Figure 23. Cumulative Distribution of Section I SMOG Readability Score

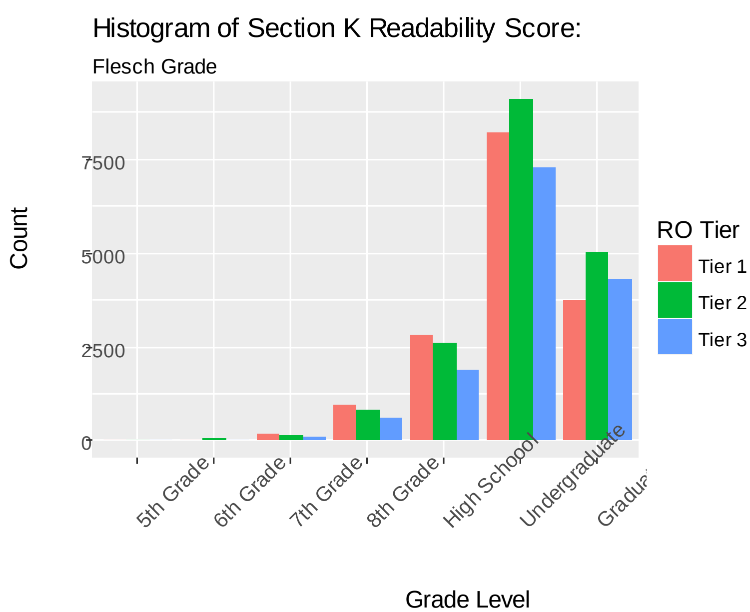


Figure 24. Histogram of Section K Flesch Grade Level

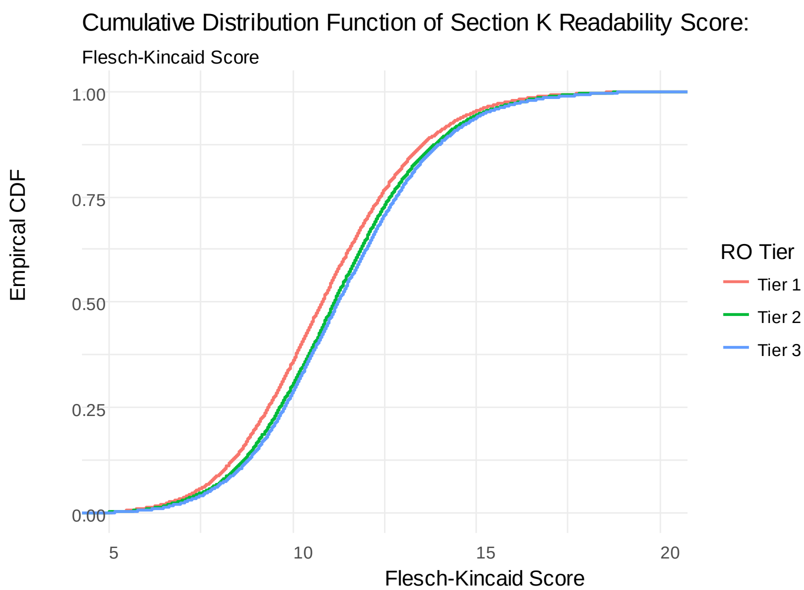


Figure 25. Cumulative Distribution of Section K Flesch-Kincaid Readability Score

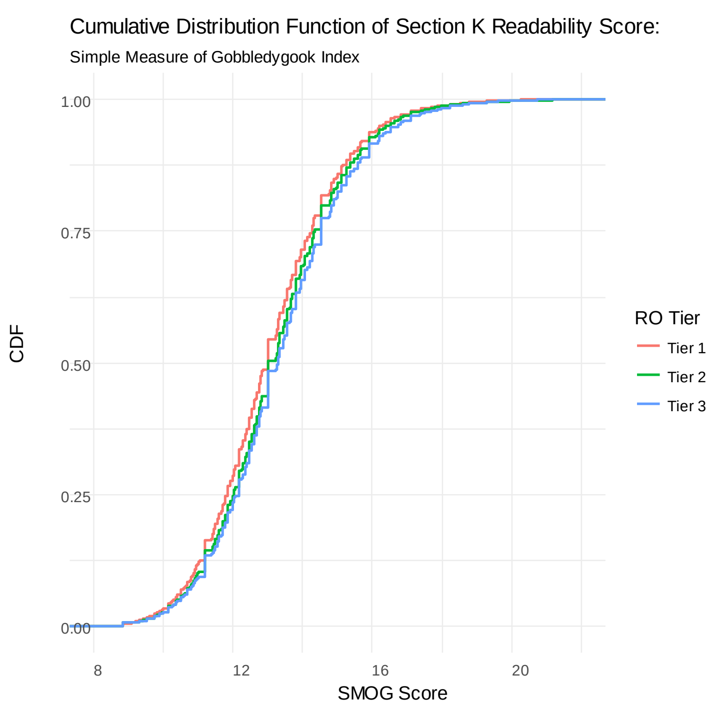


Figure 26. Cumulative Distribution of Section K SMOG Readability Score

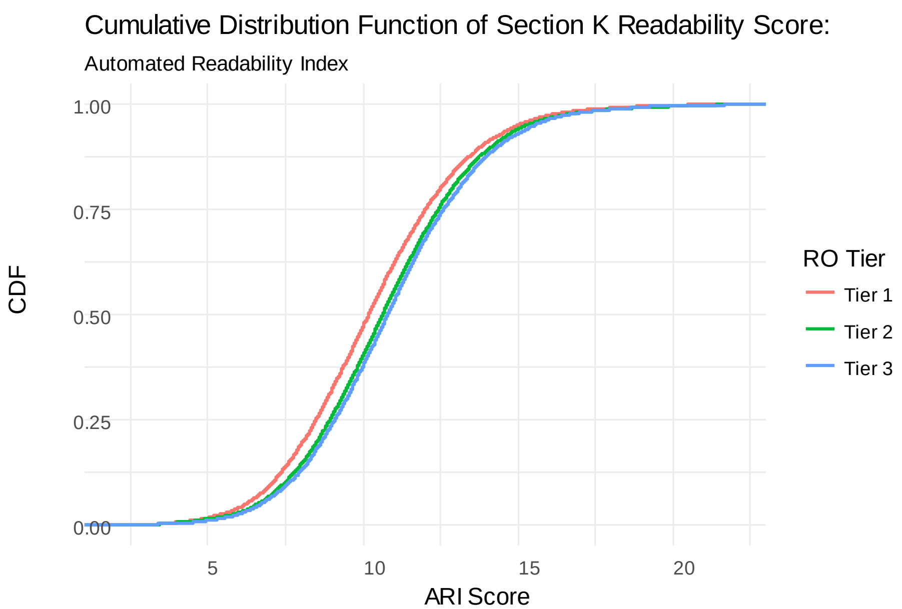


Figure 27. Cumulative Distribution of Section K ARI Readability Score

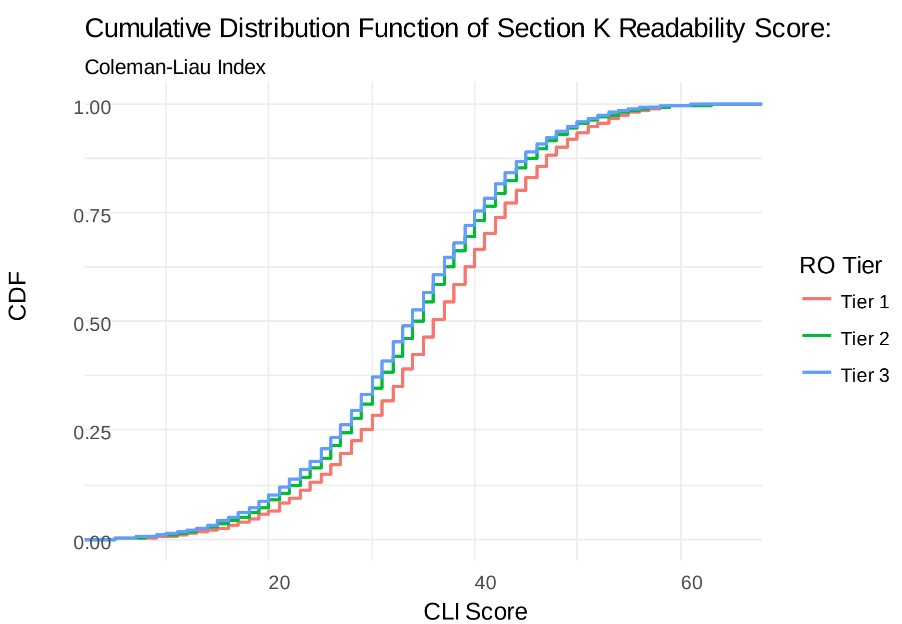


Figure 28. Cumulative Distribution of Section K Coleman Liau Readability Score

We use the Jonckheere Terpstra test to determine whether the trending separations between each tier group occur by chance and find that each of them are statistically significant in Table 9.

Table 9. P-values for the Jonckheere and Terpstra Test on Section I and K Readability Statistics

	Sample Size				Spelling Errors	Flesch Grade	Flesch Kincaid Score	ARI	SMOG index	Coleman Liau Score
Section I	Total	1	2	3	$3.1x10^{-14}$	0.001	$4.11x10^{-6}$	$2.4x10^{-10}$	$1.3x10^{-6}$	$2.2x10^{-16}$
	45303	15906	17901	11496						
Section K	Total	1	2	3	$2.2x10^{-16}$	$2.2x10^{-16}$	$2.2x10^{-16}$	$2.2x10^{-16}$	$2.2x10^{-16}$	$2.2x10^{-16}$
	47968	15928	17809	14231						

However significant these p-values are, we find that they are not practical as the large sample size affects the quality of the test statistic. When analyzing the central tendencies of each readability statistic and observing their estimated cumulative distribution function, we find that they provide little additional value in predicting tier performance.

d. Predictive Modeling of Corpus Descriptive Statistics

Pursuing conclusions by Ghose and Ipeirotis (2011) and Jordan (2011) on the value of word counts, character counts, number of spelling errors, and five readability statistics, we use penalty inducing models to predict performance tiers. We establish the comparative base line as a naïve distribution between the three tiers. Any performance greater than 1/3 would be an improvement to our model. We train each model on the training set and test it against the validation set. Our predictive variables are word counts, character counts, number of spelling errors, Flesch Kincaid Score, Flesch Grade Level, SMOG Index, Coleman Liau Index, and the ARI, and our response variable is the performance tier. We run the model for the RS and RO.

(1) Elastic net

Using the Lasso and Elastic-Net Regularized Generalized Linear Models under **glmnet** (Friedman, Hastie, Simon, & Tibshirani, 2016), results for the tier prediction proved slightly better than a naïve assignment. We observe in Table 10 the cross validated model's performance. The model does not classify the tiers with great accuracy. The confusion matrix indicates that the model does not predict the top tier well and has a lot of overlap with tier 2. Tier 2 predicts slightly better, and as expected bleeds-over with tier 1 assignment. The most interesting observation is that the majority of tier 3 Marines are assigned to tier 2, demonstrating that only observing word statistics without their substance, does not lead to consistently predict tier 3 Marines.

Table 10. Classification Performance of Generalized Linear Model on Corpus Descriptive Statistics

		Prediction								
		Reporting Senior					Reviewing Officer			
		Tier 1	Tier 2	Tier 3	Total		Tier 1	Tier 2	Tier 3	Total
Actual	Tier 1	1,490 (38%)	2,240 (57%)	205 (5%)	3,935		1,990 (50%)	1,603 (41%)	349 (9%)	3,942
	Tier 2	1,272 (29%)	2,759 (62%)	421 (9%)	4,452		1,574 (35%)	2,188 (49%)	707 (16%)	4,469
	Tier 3	543 (19%)	1,817 (64%)	483 (17%)	2,843		1,040 (27%)	972 (26%)	1,781 (47%)	3,793
	Total	3,305	6,816	1,109	11,230		4,604	4,763	2,837	12,204

We calculate the correct classification rate by

$$\frac{TrueNegative + TruePositive}{Total}$$

and compare it to naïve tier assignment. The RS and RO score 43% and 42% which is better than 33.3%. Although not practically significant on its own, the general linear model provides information value to tier assignments.

We pursue how the variables changed by examining the effects of relaxing the penalty and notice that for both RS and RO, the first variable to enter the model are the

6th, 5th, and 7th grade Flesch readability levels followed by character counts. We extract the coefficients from the **glmnet** object with λ = 0.0032645 to write an equation to estimate the tiers:

$$
\begin{aligned}
Tier1 &= -1.54569 - 0.00428(\text{SpellingErrors}) - 0.07173(SMOG) + 0.09123(FleschKincaid) \\
&+0.01373(ColemanLiau) - 0.03227(\text{ARI}) - 0.00105(WordCount) + 0.00219(CharacterCount) \\
&-0.03600(FleschGradeCollege) - 0.00186(\text{FleschGradeGraduate}) + 1.59090(\text{FleschGrade 6 th}) \\
&+0.62370(\text{FleschGrade 7 th}) + 0.08460(\text{FleschGrade 8} - \text{9 th}) \\
Tier2 &= 0.06830 + 0.00916(\text{SpellingErrors}) + 0.01848(SMOG) - 0.02968(FleschKincaid) \\
&-0.00112(ColemanLiau) + 0.00623(\text{ARI}) - 0.00418(WordCount) + 0.00010(CharacterCount) \\
&+0.01684(FleschGradeCollege) + 0.05330(\text{FleschGradeGraduate}) + 0.67952(\text{FleschGrade 6 th}) \\
&-0.10584(\text{FleschGrade 7 th}) + 0.01067(\text{FleschGrade 8} - \text{9 th}) \\
Tier3 &= 1.47739 - 0.00488(\text{SpellingErrors}) + 0.05324(SMOG) - 0.06155(FleschKincaid) \\
&-0.01261(ColemanLiau) + 0.02604(\text{ARI}) + 0.00524(WordCount) - 0.00319(CharacterCount) \\
&+0.01915(FleschGradeCollege) - 0.05144(\text{FleschGradeGraduate}) - 2.27041(\text{FleschGrade 6 th}) \\
&-0.51786(\text{FleschGrade 7 th}) - 0.09486(\text{FleschGrade 8} - \text{9 th})
\end{aligned}
$$

The equation to estimate performance tiers for RO is extracted from **glmnet** with λ= 0.002874398:

$$
\begin{aligned}
Tier1 &= -1.475075 + 0.077897(\text{SpellingErrors}) - 0.083082(SMOG) + 0.120329(FleschKincaid) \\
&+0.015058(ColemanLiau) - 0.062470(\text{ARI}) - 0.000238(WordCount) + 0.003369(CharacterCount) \\
&-0.066993(FleschGradeCollege) - 0.017908(\text{FleschGradeGraduate}) + 0.226399(FleschGrade5th) \\
&-0.239819(\text{FleschGrade 6 th}) + 0.204778(\text{FleschGrade 7 th}) + 0.041981(\text{FleschGrade 8} - \text{9 th}) \\
Tier2 &= -6.60586 - .001060(\text{SpellingErrors}) - 0.008624(SMOG) - 0.015746(FleschKincaid) \\
&+0.006230(ColemanLiau) + 0.016300(\text{ARI}) - 0.000081(WordCount) - 0.000395(CharacterCount) \\
&-0.019980(FleschGradeCollege) - 0.052600(\text{FleschGradeGraduate}) - 0.275635(FleschGrade5th) \\
&+0.990338(\text{FleschGrade 6 th}) - 0.060385(\text{FleschGrade 7 th}) + 0.007439(\text{FleschGrade 8} - \text{9 th}) \\
Tier3 &= 0.814491 - 0.076837(\text{SpellingErrors}) + 0.091705(SMOG) - 0.104582(FleschKincaid) \\
&-0.009829(ColemanLiau) + 0.046070(\text{ARI}) + 0.0003188(WordCount) - 0.002974(CharacterCount) \\
&+0.086974(FleschGradeCollege) + 0.070509(\text{FleschGradeGraduate}) - 0.502033(FleschGrade5th) \\
&-0.750519(\text{FleschGrade 6 th}) - 0.144393(\text{FleschGrade 7 th}) - 0.049417(\text{FleschGrade 8} - \text{9 th})
\end{aligned}
$$

The generalized linear model provides insight on the value of the Flesch Grade level readability statistics. By examining the tier-estimation equations, we visualize the

magnitude each predictor affects the model. We see that while analysis in Chapter V. Section A.2 show the greatest separation between tiers are found with word and character counts, they provide marginal value in the generalized linear model. Although this elastic net does not have strong predictive power, it provides value in an ensemble.

(2) Classification and Regression Trees

We use the classification and regression tree as an alternative machine learning technique to gain more insight on valuable information from corpus statistics. Table 11 represents the classification performance of the cross validated model with the marginal distribution of actual tiers. We observe that the model performs well when classifying the middle tier, but not for the top or bottom tiers. We calculate the correct classification rate by

$$\frac{TrueNegative + TruePositive}{Total}$$

and compare it to naïve tier assignment. The RS and RO score 41% and 43%, which is better than the 33.3% naïve assignment.

Table 11. Classification Performance of Generalized Linear Model on Corpus Descriptive Statistics

		Prediction							
		Reporting Senior				Reviewing Officer			
		Tier 1	Tier 2	Tier 3	Total	Tier 1	Tier 2	Tier 3	Total
Actual	Tier 1	2,006 (51%)	1,700 (43%)	232 (6%)	3,938	1,462 (37%)	1,498 (38%)	982 (25%)	3,942
	Tier 2	182 (6%)	2,209 (78%)	428 (15%)	2,819	1,110 (25%)	1,717 (38%)	1,642 (37%)	4,469
	Tier 3	810 (28%)	1,550 (54%)	485 (17%)	2,845	671 (19%)	1,250 (35%)	1,628 (46%)	3,549
	Total	2,998	5,459	1,145	9,602	3,243	4,465	4,252	11,960

After cross validating and tuning the model in **rpart** (Therneau, Atkinson, & Ripley, 2015) to optimal conditions, we plot the tree using **rattle** (Williams, 2011).

Figure 29 provides a guide of how to interpret the tree plots. For each split, the prediction becomes the maximum value of predicted tiers. In this case, 0.39 of the documents were classified as tier 2, which is larger than all the others; therefore, in this split, all documents are classified as tier 2.

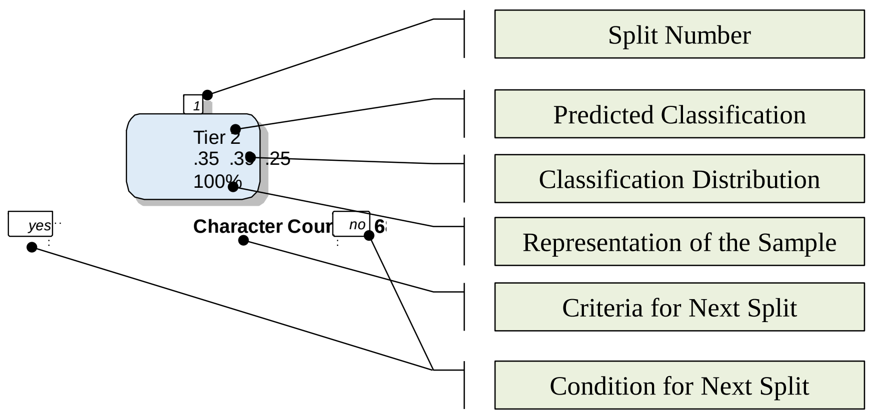


Figure 29. Explanation of **rattle** Output Tree

In Figures 30 and 31, we examine the RS and RO classification regression tree and observe that the model uses different predictive variables than the elastic net to achieve comparable results.

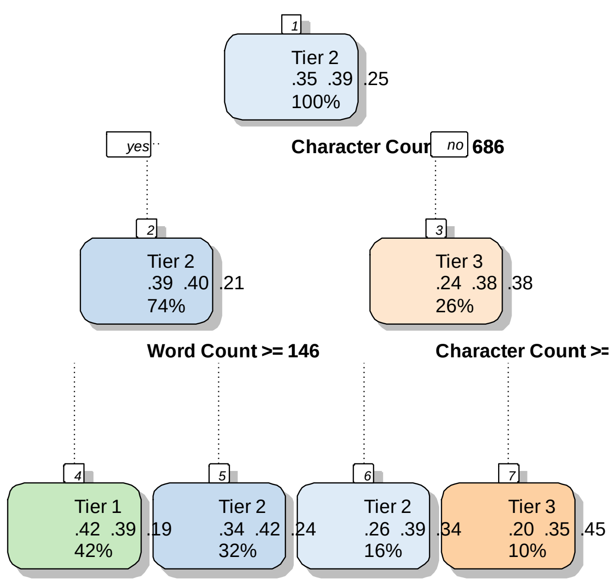


Figure 30. Section I Classification and Regression Tree

In this case of Section I, comments with more than 686 characters and 146 words are classified tier one. Alternatively, comments with less than 546 characters are classified as tier 3. The rest are classified as Tier 2. The RO tree uses word counts above 90 and ARI score below 8.4 to predict tier 1 MROs. Tier 3 Marines are only classified when the word count is less than 70.

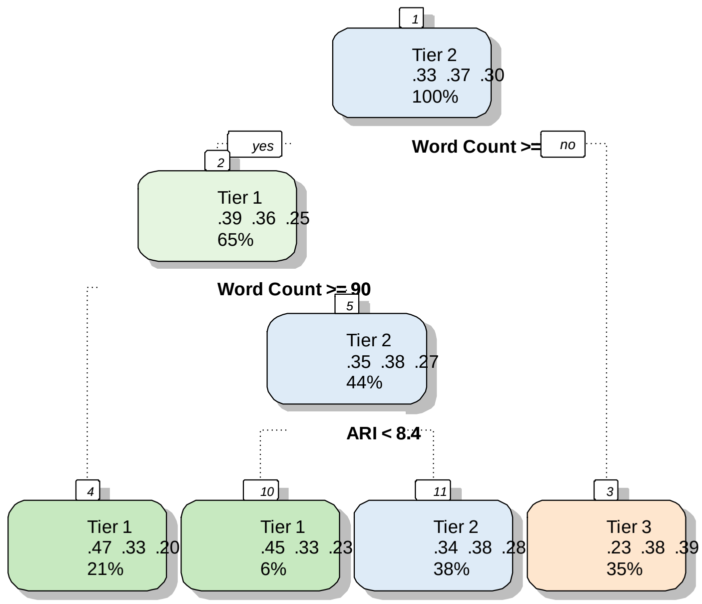


Figure 31. Section K Classification and Regression Tree

We determine that the Flesch Grade level provides additional predictive value when using the generalized linear model and word counts, ARI score, and character counts when using classification and regression trees. Each model had a tendency of predicting the middle well but for different reasons. We will leverage the strengths of these two models in Chapter V, Section C.2, when we build a stacked ensemble of models.

3. Directed Comments

We investigate whether RSs comply with the PES manual by considering how directed comments are implemented in the textual analysis. The motivation behind this approach is that directed comments provide a list of items that need to be addressed aside

from the word picture. By addressing specific and important information, it allows the reporting senior to focus on the intangibles of the word picture. The PES manual explicitly directs that RSs "begin each directed comment with the entry 'Directed Comment' and a reference to its origination in the report (e.g. 'Directed Comment. Sect A, Item 6a:')" (Commandant of the Marine Corps, 2015, p. 4-39).

We consider five different scenarios: observation time less than 90 days, operational risk management, pilots, duty in combat zone, and awards. We use these cases because they each require a specific directed comment that is easily greppable. When a FitRep observation period is longer than 30 days but less than 90 days, the RS has the option to write an observed report and must comment on how he or she believes they have enough observation time to evaluate the MRO's performance. In the data set, there are 4,146 fitness reports with less than 90 days of observation time. Of those 4,146 fitness reports, only 1,545 RSs have a directed comment and of those, only 555 or 13.4% have a directed comment one specifying the reason for the short report.

By design, Marine Officer billets involve the ORM principles of "planning, supervision, training, and operational responsibilities" (Commandant of the Marine Corps, 2015 p. 4–43). In this case, we expect that every officer has a fitness report that targets one of those duties and executes ORM. Out of the 50,267 officers, only 17,933 have a directed comment, and of those, only 1,177 or 2.34% comment on an officer's compliance with the ORM principles.

By considering billet military occupational specialty codes (BMOS) and designated units, we screen pilots that are assigned to operational units. Of 50,267 screened observed fitness reports, we isolate 12,015 FitReps mapped to pilots in a flying billet. Based on the directed comment "in the case of Marine aviators and flight officers, comment on pure flying proficiency" (Commandant of the Marine Corps, 2015, p. 4-43), we expect those 12,015 pilots have a comment on their flying proficiency. When grepping for directed comments and words related to flying, we only find 3,925 with directed comments, of which, 285 or 2.37% pertain to flying. After further investigation, comments towards flying proficiency are traditionally captured in section C. Although a departure from the PES manual, a pilot's proficiency is not lost information.

Contrary to the three cases above, Marines who are subject to commendatory material are well represented. Of the 50,267 FitReps in the dataset, 8,689 are commendatory fitness reports. Of that subset, 7,053 or 81.2% have a directed comment on the commendatory nature of the marking. The high number of completion could be as a result of the automatic prompt window that occurs when the RS checks the commendatory box in section a, item 6.a.

Finally, we look at whether a directed comment is listed when a Marine serves in a combat zone. When this instance occurs, the RS checks section A, item 3.c of the fitness report offers the option to select "C" for combat zone. The RS is supposed to "comment on the nature of the combat operation and the MRO's actions related to the operation" (Commandant of the Marine Corps, 2015, p. 4-40). We find that of the 8,306 observed combat fitness report, 5,158 or 62.1% of fitness reports had the requisite directed comment. Similar to checking the commendatory block, the combat zone block provides a prompt to remind the RS to comment on the nature of the fitness report.

Although investigating each of the 41 instances a directed comment is applicable would be time difficult, by observing these five common case studies, we discover that the majority of reporting seniors do not comply with the directed comment guidance in the PES Manual. When we trace the educational material, FitRep handbook, and lesson plans from the Basic Officer Course at TBS, we see that these documents do not dedicate more than a sentence on the directed comments (The Basic School, 2016c, p. 2; Dodd, 2016, p. 30). This discrepancy might be attributed to the lack of education or residual knowledge from the old PES. Since the RO is required to check and comment on whether the fitness report is "administratively correct" (Commandant of the Marine Corps, 2015, p. 8–2) before forwarding the fitness report to Headquarters Marine Corps, an implemented fix is educating the RS and RO on the purpose, value, and implementation of the directed comments.

4. "Number 1" Analysis

The previous PES manual (NAVMAC 2794) encourages ranking the MROs among his peers within the reporting period (United States Marine Corps, 1995, Chap 6)

potentially due to the lack of ability to mass compute relative values and comparative assessments in the old PES. Example would be “ranked #1 of 24 Capts in the unit” or “Top Major in the Department.” The new PES manual does not require to make a quantitative statement in the mandatory comments because APES automatically calculates the appropriate score at the time of processing and cumulatively. Perhaps due to residual knowledge from the old PES manual or a lack of eduation, RSs and ROs routinely continue to annotate the Marine’s relative placement within his peer group in the comment fields.

Within the current structure of the APES, we can juxtapose the relative value with the in-comment ranking of MROs. We use regular expressions to separate instences in Sections I and K when the RS or RO utilize terms as “Number one,” “Best [rank]…,” “Top [Rank] in …,” “#1 of [n],” amongst others. When an RS grades a MRO with a top score, his relative value becomes 100%. We can only examine those at the top due to the high variablity of relative values other than the best and worst MRO. Conversely, the RO does not have a score that can easily distinguish the MRO from his peers; however, we can examine how many are above, with, or below the MRO. If the Marine was truly the best, he or she would have no one above, a few people with, and a lot of people below him or her.

We observe in Table 12 that given the MRO is mentioned as being the best, the RS only marks him the best 37.5% of the time at processing. The relative values of those getting the highest endorsement at processing are spread across the three tiers due to the presence of low density profiles. As RSs gain more observations, the tier representation gradually shifts towards the top third. We notice the decrease observations between at processing and cumulative which is expected as MROs are unseated as “the number one.”

Table 12. MRO Distribution of RS Tiers Based on “Number One” Comments

	100	Tier 1	Tier 2	Tier 3	No Profile
% Cumulative	27.8%	62.8%	23.7%	5.4%	8.1%
% at Processing	37.5%	44.0%	15.5%	32.2%	8.3%

The RO's comments show a trend that is consistent with the comparative assessment distribution analysis in Chapter V, Section A.1: top MROs are not separated from the preponderance of their peers. While only 7% are actually above the value, 61% are labeled in the same block as the MRO, and 32% are below. Using Majors and Lieutenant Colonels as examples, we see that the majority of observations are contained in the top blocks. Comments towards being the top Marine would not quantifiably distinguish the individual from others within the same evaluation block.

Labeling an officer with superlatives akin to being "number one" within his peer group does not provide informational value on its own. While seeing a "100" FitRep leads an MRO to the conclusion that he or she is "number one," the opposite is not true due to inconsistencies with the assigned marks.

B. KEYWORD ASSOCIATION

To determine the commonly used words associated with each tier within a particular grade, we focus our attention on specific comments related to "potential for promotion and assignments to command, staff, and advanced schooling" (Commandant of the Marine Corps, 2015, p. 4-19). We take the unigram token matrix of each rank split into the three tiers and apply three different word association models: a supervised correlation, unsupervised correlation, and syntagmatic. We define supervised keyword correlation as finding the terms that are correlated with a target keyword. Derived from the Commandant's guidance, these words are "promote," "retain," "potential," and "assign." We add the words "command" and words associated with professional military education (e.g., career level School (CLS), Intermediate level school (ILS), top level school (TLS)). For any schooling options, we use regular expressions to account for the use of acronyms, different tracks such as the Maneuver Captains Career Course or Expeditionary Warfare School, and appropriate grade representation. In the model, these words are stemmed to ensure minor variations of the word due to grammatical use or misspelling does not influence the intent of the word. Alternatively, unsupervised keyword correlation examines the entire corpus for correlated term without our guidance. Syntagmatic relations are defined as correlated occurrences where certain

words within close proximity to others tend to consistently occur (Zhai, 2017). As recommended by the PES manual, we search for : promote, retain, assign, The end state is to determine whether there is a pattern between ranks and tiers of key descriptors of these key terms.

1. Supervised Keyword Correlation

The weighted term-document matrices we use are stemmed to allow for grammatical variations of the words, but assuming a consistent sentiment to the word. We use the FindAssociation () function in the **tm** package to extract the Pearson correlation coefficients and then we rank each term based on the magnitude of the number. We only take terms that are positively correlated with each other into consideration. Positive correlation implies that seeing a word with a key term is an indicator for that tier. A negative correlation would imply that the absence of term indicates a tier. We assume that a board member reading a specific word complementing a key term is more powerful than having to interpret what the RS was implying by not saying something. We set the threshold at the lower bound of weak correlation (0.20) because the terms are generally weakly correlated. Appendix C details each of the key words by rank and by term. If a specific field figure is blank, it is due to having no words at least weakly correlated with the key term.

Generally, we find that there is limited consistency in correlations between descriptive words and key terms. When overlaying tier groups on a specific graph we notice certain trends: (1) for education, potential, and promotion, the correlations tend to converge on similar terms; (2) comments on future potential of command are reserved for top tier groups; (3) comments on general assignment and retention are weakly correlated or uncorrelated. We provide some examples below to illustrate the trend, but each graph is represented in Appendix C.

Words associated with education, potential, and promotion generally converge to the same terms. Figure 32, illustrates that for the word “potential” used in captains’ section I comments, “unlimited” and “growth” are associated with all tier levels. Although the magnitude of the correlation coefficient is slightly different, the terms are

ranked the same and there are no alternatives above weak correlation. This trend is similar for education and promotion for each rank. The lack of variety by tier for associated word combined with the tendency of words to converge on the same terms renders comments on education, potential, and promotion to add little value on predicting an officer's tier.

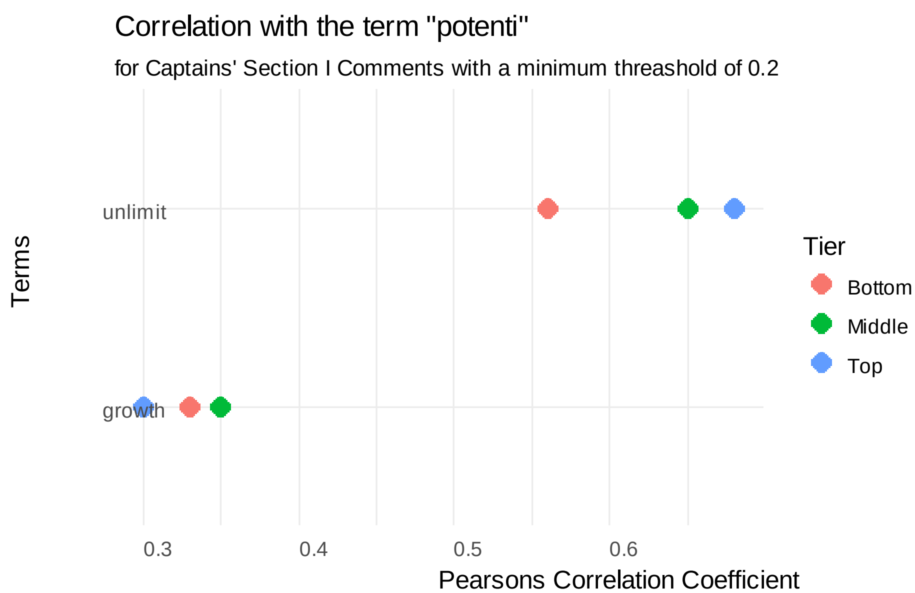


Figure 32. Correlation with term "potential" for Captains' Section I Comments

When observing word correlations, the most powerful indicators of tiers are comments directed towards command. Comments related to the MRO's current command level are for bottom tiered Marines. Conversely, potential for future command opportunities, or command at a level beyond the expected grade level, are ordered from most common to least common for top tiered Marines. For example, a 1stLt would be occupying the billet of platoon commander and the next level would be company command, which is generally occupied by a captain. Figure 33 demonstrates that each tier is ordered based on the current command from top to bottom tier and future command potential from bottom to top tier. This theme is consistent with captains occupying company command billets and their potential to assume battalion command.

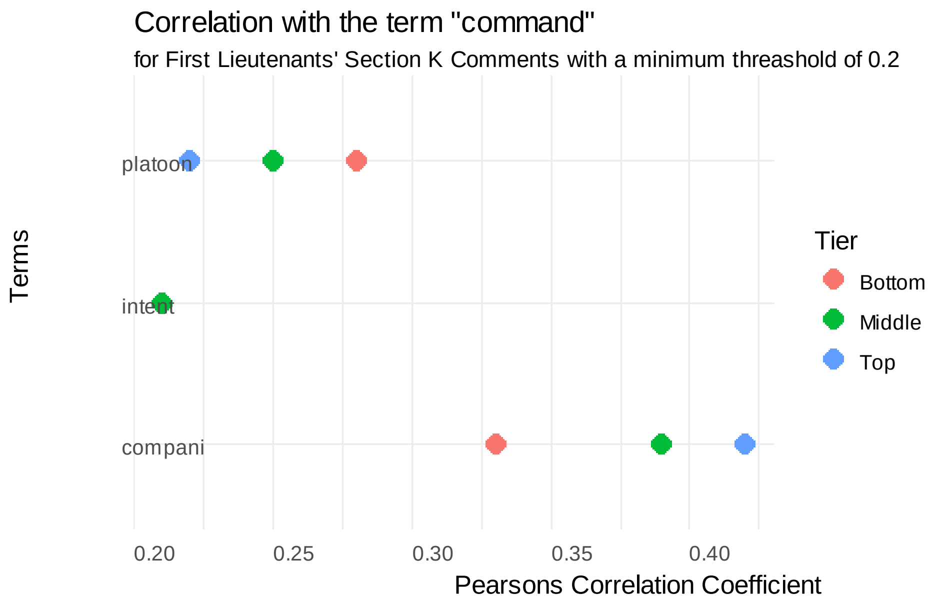


Figure 33. Correlation with term “command” for 1stLts’ Section K Comments

When the word “peer” is associated with any key term, it is generally reserved for bottom tier Marines. Figure 34 provides an example for second lieutenants.

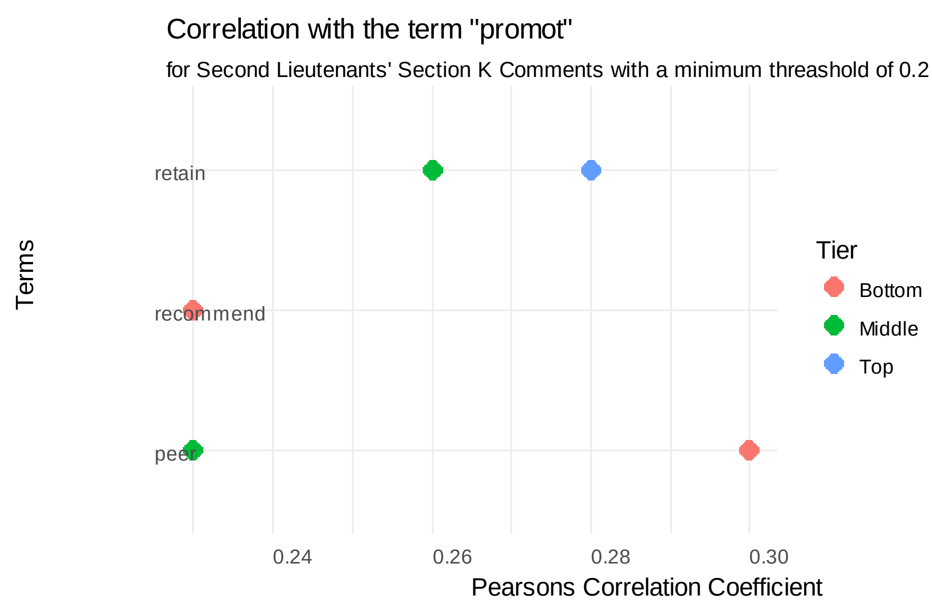


Figure 34. Correlation with term “promote” for 2ndLts’ Section K Comments

Retention and general assignment consistently yields correlations below 0.20 which is expected because they are not explicitly required. The above mentioned trends are consistent for each rank and between RSs and ROs.

2. Unsupervised Keyword Correlation

Similar to the supervised model, we use stemmed, weighted unigram term-document matrices for each rank and each tier. Figure 35 illustrates the results of this analysis. The function outputs a correlation map with terms that occur over a minimum frequency circled in red and lines that connect correlated terms above a minimum correlation threshold. The minimum frequency and correlation threshold are inputs to the function and provided in the title of the plot. The thicker the black line, the more correlated the term. While these graphs do not produce a correlation coefficient, they provide insight by displaying visual patterns that enhance discovery of potential relationships we missed in our supervised correlation model. Generally, the most correlated terms are associated with administrative requirements. Because we did not target specific terms, words like “directed comment” and “operational units” or “combat” are correlated throughout the corpus.

We find the few common trends are consistent with the supervised learning model used in the prior section, and we did not discover any unanticipated relationships. For example, Figures 35 and 36 detail top and bottom captains’ corpuses and demonstrate that comments for command are different by tier and comments on education, promotion, or potential are comparable between tiers. These trends are consistent across ranks and further illustrations are available in Appendix B.

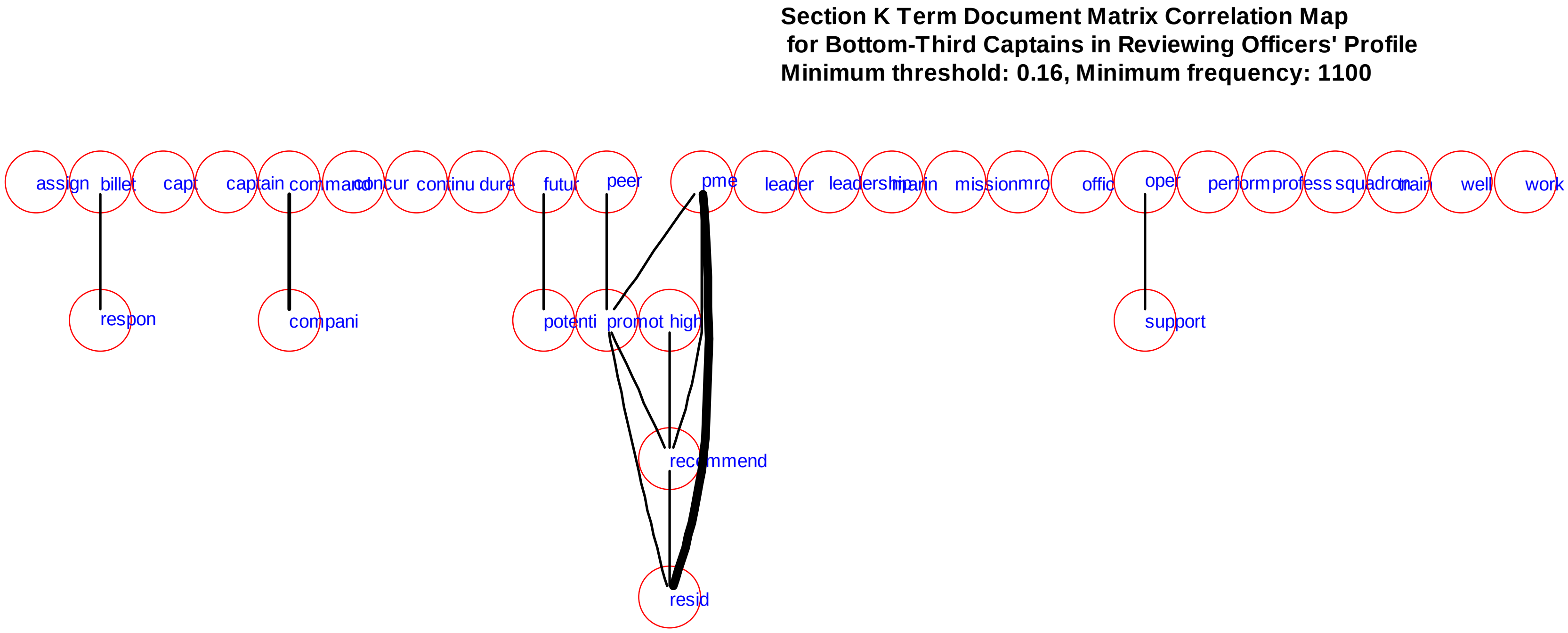


Figure 35. Section K Term Document Matrix Map for Bottom Third Captains

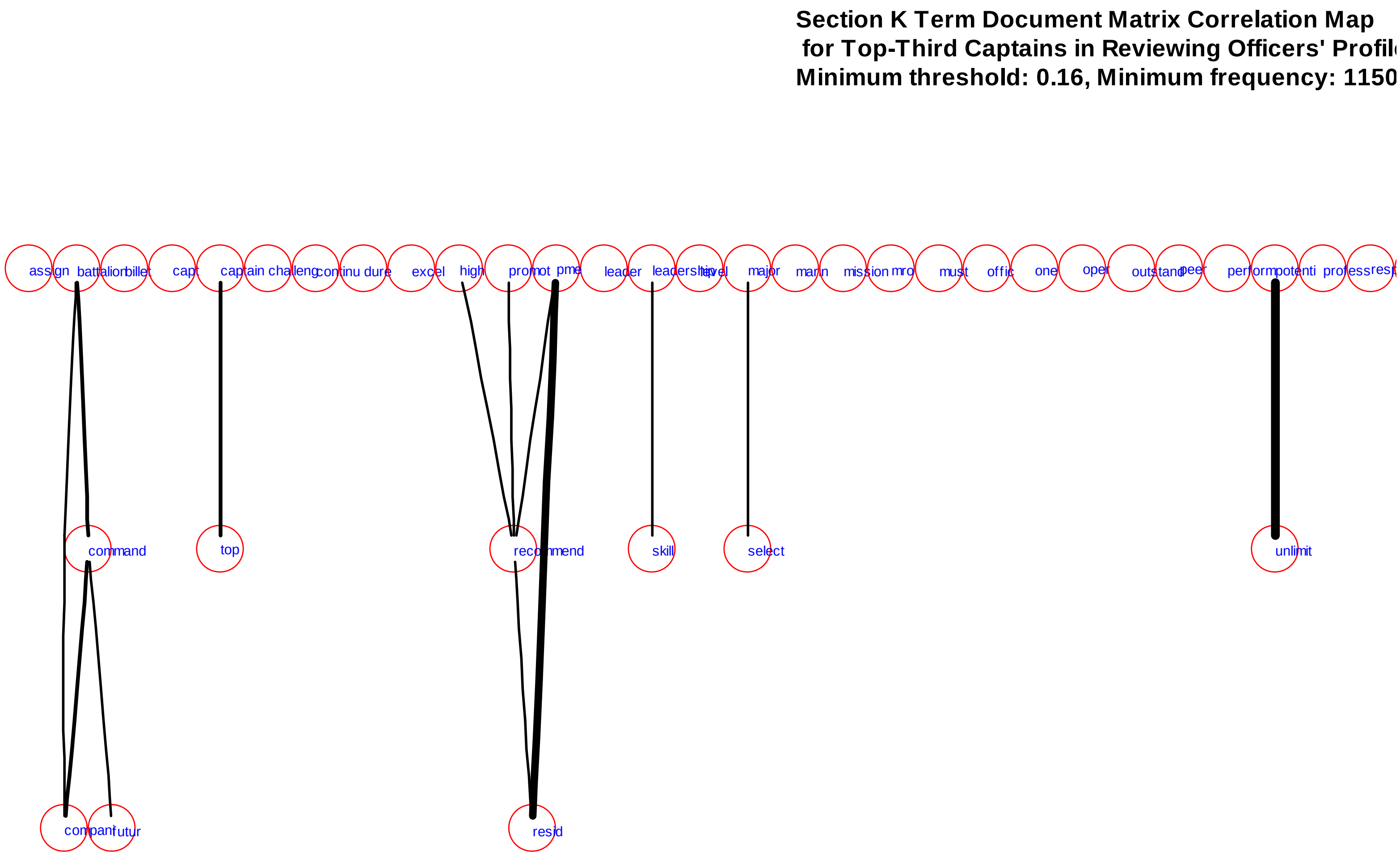


Figure 36. Section K Term Document Matrix Map for Top Third Captains

When observing all the unsupervised correlation plots in Appendix C, we observed section K comments generally have less variability in terms and relationships. Contrary to all the other ranks, second lieutenants have so much variability in words and correlations in their fitness report corpus that insight on keyword relationships was difficult to visually interpret.

3. Syntagmatic Word Association

We use an alternative approach to finding relationships between words by examining syntagmatic word association. Syntagmatic relationships can be further defined as “A & B have syntagmatic relation if they can be combined with each other” (Zhai, 2017) and represent a consistent idea. We use **text2vec** (Selivanov, text2vec, 2016) as an alternative technique to represent the corpus. As a departure from the use of bag-of-words to store corpus information, Selivanov (2016) stores words in a single vector. For each word in a corpus, the software stores the user-inputted *m*-words prior to and after each word (*n*) as a separate row and simply counts occurrences. The information is then stored as a sparse matrix which reduces computation time significantly (Selivanov, text2vec, 2016).

Using this vector of words to *n*-terms, we train a neural network and reduce it from $n \times m!$ dimensional space, where n is the number of words and m is the number of terms left and right of the word, to 50. We take our key terms as the response variable and back trace through the hidden layer looking for which words have the highest coefficient based on cosine similarity. This cosine similarity index is found by

$$\frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2 \sqrt{\sum_{i=1}^{n} b_i^2}}}$$

where **a** and **b** are vectors of the number word occurrences with components a_i and b_i.

Similar to the other models, the results from the unsupervised correlation model were not helpful in finding words associated with promotion, retention, and education.

We find the same relationships between tiers for the key terms “command” and “promote” for captains and lieutenant colonels. While most of the observations are not helpful, we notice a couple of trends based on our results pictured in Tables 13 through 16. For example, superlatives like “best” are in all three tier groups. As a departure from our previous observations, comments based on “peers” now appear in all three tiers.

Table 13. Most Common Terms Syntagmatically Associated with “Command” for Lieutenant Colonels

top tier words	top tier coef	middle tier words	middle tier coef	bottom tier words	bottom tier coef
tenacious	0.4321	tour	0.5296	tour.	0.5241
Operations_Officer.	0.4297	committed_to	0.4499	the_best	0.4614
operational	0.4261	in_joint	0.4442	the	0.4391
had	0.4238	recommendation_selection	0.4184	RS,	0.4096
same.	0.4228	guidance.	0.4117	Marine	0.4053
to_achieve	0.4146	Officer_the	0.4087	He's	0.3896
flawlessly	0.4144	she	0.4034	TLS	0.3800
School.	0.4101	Promote_first	0.4014	_LtCol	0.3785
RS.	0.4074	standard	0.3974	officer_who	0.3737
has_already	0.4051	has_contributed	0.3767	one_the_most	0.3720
with_RS._	0.4035	the_planning	0.3753	the_example	0.3650
the_best	0.3994	MAG-14	0.3734	MARSOC.	0.3577
my_top	0.3944	Bravo	0.3696	select	0.3558
selected_to	0.3901	the_readiness	0.3613	_Will	0.3558
peers._	0.3768	grade._	0.3561	best	0.3546
standards.	0.3620	subject	0.3503	I_concur_with	0.3498
pleasure_to	0.3616	in_short	0.3486	TLS.	0.3474
directly_to	0.3582	short_reporting	0.3462	the_RS	0.3455
his_peers.	0.3579	Advocate	0.3417	peers,	0.3442
as_Battalion	0.3506	constantly	0.3411	team_player	0.3409

Table 14. Most Common Terms Syntagmatically Associated with “Command” for Captains

top tier words	top tier coef	middle tier words	middle tier coef	bottom tier words	bottom tier coef
potential_&	0.5912	presence_in_the	0.5377	the_battalion	0.4760
opportunities.	0.5391	the_MAWTS-1	0.5133	lapse_in_judgment	0.4661
period._MRO	0.4655	Vast	0.4951	results_in_the	0.4505
company	0.4584	his_role_as	0.4937	great_asset	0.4454
_Also	0.4496	MAWTS-1	0.4842	opportunities.	0.4442
_Successfully	0.4486	control	0.4834	squadron_in_the	0.4438
Marine_Officer_who	0.4457	year	0.4732	on_to_make	0.4333
Landing	0.4432	the_mark.	0.4723	squadron_to	0.4306
to_positively	0.4355	company	0.4715	battalion	0.4294
impressive	0.4267	intent,	0.4692	subordinates_alike._	0.4262
&	0.4249	attributable	0.4639	will_make_an	0.4100
future_promotion.	0.4184	on._	0.4629	track_to	0.4089
have_lasting	0.4175	AO_Denver._	0.4624	exposure	0.4081
pilot._	0.4173	himself	0.4584	offered	0.4078
&_his	0.4173	with._A	0.4518	to_the_unit.	0.4060
expertise_in	0.4151	in_the_middle	0.4508	Palms._	0.4060
pilot.	0.4139	immersed	0.4491	demanding_billet	0.4051
Culver	0.4118	professional._+	0.4469	advise	0.4050
superbly.	0.4096	counsel._	0.4388	Alen	0.4014
future	0.4042	Denver.	0.4381	_I_enthusiastically	0.4005

Our results provided in Tables 15 and 16 show that other than a favorable comment of “most enthusiastic recommendation,” the path towards tier prediction through the neural network’s hidden layers is inconclusive.

Table 15. Most Common Terms Syntagmatically Associated with “Promote” for Lieutenant Colonels

top tier words	top tier coef	middle tier words	middle tier coef	bottom tier words	bottom tier coef
SNO_has	0.4736	coach	0.4778	forces.	0.4259
support_to_the	0.4480	professionalism.	0.4712	front	0.4140
Marines_are	0.4445	critical_in	0.4331	increase	0.3978
assuming	0.4409	incredible	0.4277	two	0.3885
Colonel	0.4331	wide	0.4092	to_come.	0.3804
resource	0.4063	deployed	0.4026	to_TLS.	0.3734
is_future	0.4053	in_blue,	0.4013	Commander.	0.3727
most_demanding	0.3939	most_complex	0.3987	tactical	0.3541
to_make	0.3905	consummate_professional	0.3950	OIC	0.3533
mission._	0.3904	knowledgeable	0.3941	Operations_Officer	0.3471
Commanders_in_3d	0.3811	operate	0.3865	impact_on_the	0.3402
most_enthusiastic_recommendation	0.3765	funding	0.3853	mission_accomplishment.	0.3338
leads_his	0.3699	period_as	0.3804	Col.	0.3336
to_have	0.3668	performed_superbly	0.3803	USMC	0.3310
to_command	0.3663	command_tour,	0.3720	Operations	0.3302
can_be_counted	0.3641	performance_leading	0.3697	RS,	0.3295
to_work	0.3598	has_done_an	0.3681	reporting_period	0.3260
is_one_the	0.3576	place	0.3670	region	0.3233
execution_the	0.3548	core	0.3663	his	0.3226
be	0.3524	RS_comments	0.3648	command_tour.	0.3222

Table 16. Most Common Terms Syntagmatically Associated with "Promote" for Captains

top tier words	top tier coef	middle tier words	middle tier coef	bottom tier words	bottom tier coef
employment_the	0.4897	Creative	0.4902	Corps_in	0.5087
second	0.4771	to_leverage	0.4840	Recommended_appropriate	0.4881
PME,_return	0.4738	diligence	0.4837	Assign_to_resident	0.4796
simultaneously.	0.4574	my_best	0.4760	or_Operations	0.4669
the_division.	0.4566	rest_the	0.4505	reporting_period._	0.4656
Operationally	0.4516	island	0.4484	expectations_in	0.4553
largest_most	0.4510	with_demanding	0.4443	workups	0.4489
to_be_an	0.4509	garrison._	0.4425	time_in_grade	0.4425
top_three	0.4477	important_role_in	0.4413	_Recommended_continued	0.4412
himself_the	0.4455	the_box	0.4387	with_very	0.4365
during_an	0.4271	Marines_with	0.4377	watching	0.4347
three_captains	0.4249	ahead_the	0.4360	Retain	0.4326
to_do	0.4247	garrison.	0.4348	with_tremendous	0.4247
MRO_is_one	0.4218	academic	0.4288	program_with	0.4247
to_make	0.4212	officer,_aviator	0.4261	combat_operations	0.4221
to_serve_as	0.4197	central_to	0.4199	with_solid	0.4212
the_mission	0.4173	executed_all	0.4178	in_managing	0.4188
as_combat	0.4171	program_to	0.4161	put	0.4179
problems,	0.4153	UH-1N_to_the	0.4155	Makes	0.4141
effectively_led	0.4126	trained,	0.4149	to_Company	0.4102

C. PREDICTIVE-MODEL RESULTS

1. Model Performance

To evaluate the performance of our models, we use the harmonic mean between precision and recall. Precision, or confidence, is the calculated by $TruePositive/(TruePositive + FalsePositive)$ (Powers, 2011, p. 39) and answers the question "how many of those identified are actually a success or failure." Recall, or sensitivity, is $TruePositive/(TruePositive + FalseNegative)$ (Powers, 2011, p. 39) and answers the question "how many successes are correctly identified." The harmonic mean is calculated by $(2 \times Precision \times Recall)/(Precision + Recall)$ and "references the True Positives to the Arithmetic Mean of Predicted Positives and Real Positives, being a constructed rate normalized to an idealized value" (Powers, 2011, p. 41).

The heat maps (**ggplot2)** shown in Figures 37 and 38 display a summary of the minimum harmonic mean for all our model configurations. The generalized linear model, maximum entropy, random forests, and support vector machines generally perform better than the neural network or the boosting model. The graphs also show that

adding additional interactions between word tokens does not improve the model performance. For complete numeric values of each model prediction, refer to Appendix E.

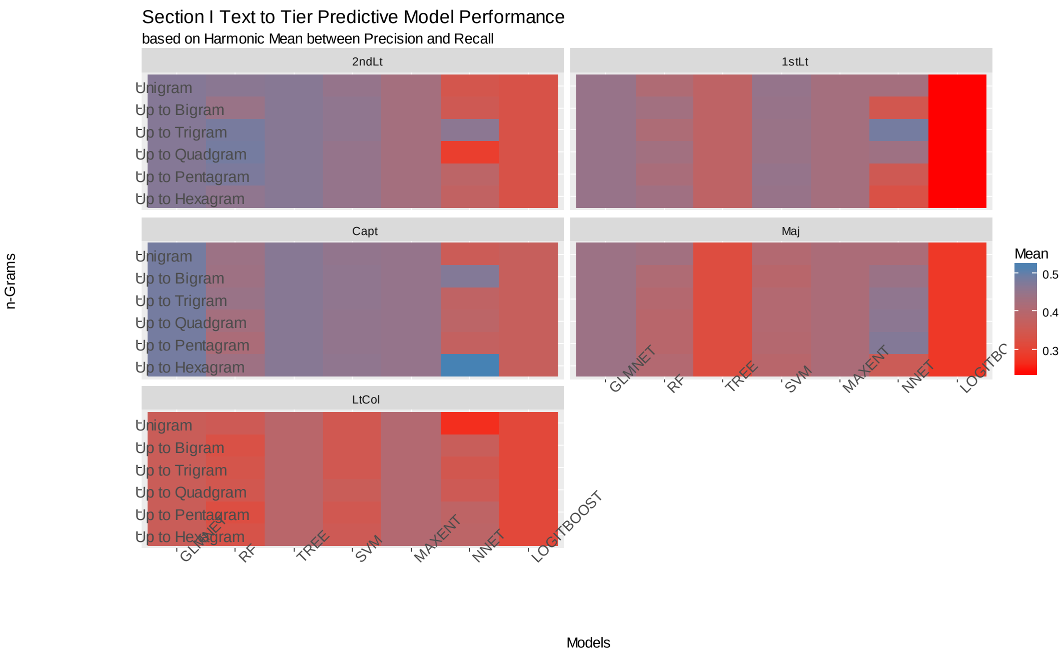


Figure 37. Section I Text to Tier Predictive Model Correct Classification Rates by Rank, by Model, by n-Gram Based on the Minimum Harmonic Mean

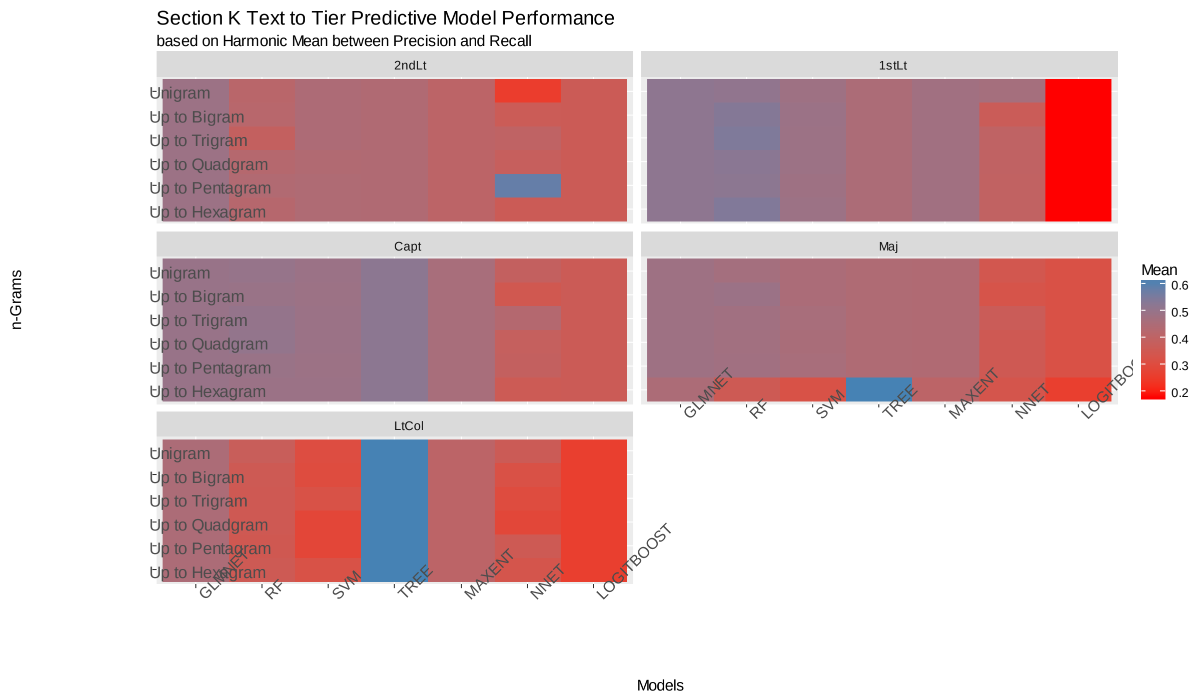


Figure 38. Section K Text to Tier Predictive Model Performance by rank, by model, by n-gram based on the Minimum Harmonic Mean

Tables 17 and 18 indicate how many models performed better than naively tier selection. These tables are a representation of the heat maps, based on count, and help identify which models have a better tendency of predicting performers. The maximum possible score is “5” due to the five ranks. We count every occurrence of a model’s minimum correct classification rate that is higher than 1/3. Since both the count and the harmonic mean show that interactions represented by multi-gram tokens do not enhance the model we only keep unigrams in the model ensemble.

Table 17. Count of Reporting Senior Models that Predicted Better than Naïve Assignment

	Glmnet	Maxent	Boosting	NNET	Tree	Forest	SVM	Total
Unigram	4	4	0	0	2	0	3	13
Up to bigrams	4	4	0	2	2	0	3	15
Up to trigrams	4	4	0	4	2	0	3	17
Up to quadgrams	4	4	0	1	2	0	3	14
Up to pentagrams	4	4	0	1	2	0	3	14
Up to hexagram	4	4	0	1	2	0	3	14
Total	24	24	0	9	12	0	18	81

Table 18. Count of Reviewing Officer Models that Predicted Better than Naïve Assignment

	Glmnet	Maxent	Boosting	NNET	Tree	Forest	SVM	Total
Unigram	4	4	0	1	3	4	4	20
Up to bigrams	4	4	0	0	3	4	4	19
Up to trigrams	4	4	0	1	3	4	4	20
Up to quadgrams	4	4	0	1	3	4	4	20
Up to pentagrams	4	4	0	1	3	4	4	20
Up to hexagram	3	3	0	0	3	3	3	15
Total	23	23	0	4	18	23	23	114

2. Ensemble

To further increase predictive power, we will use the best models outlined above and add the pre-corpus predictions to create an ensemble. The five best corpus predictive models are elastic net, maxent, support vector machines, trees, and random forest with up to three word tokens. Although not a high performing model, we use classification and regression trees as an ensemble is used as a tie breaker due to the even number of models. By adding this tie breaker, we improved our correct classification rate by 17%. Both the

tree and the elastic net pre-corpus predictive have better predictive power than the corpus models so we use both. We integrate these models into a data set with each of the model prediction labels evaluated against the validation set. Finally, we run the lasso regularized generalized linear models with the actual labels against the test set (Hastie et al., 2009, 609). This additional layer is used to find the combination of models that achieve the best results. Our results for each ensemble technique are displayed in Table 19.

Table 19. Ensemble Predictive Model Correct-Classification Rate

	2ndLt	1stLt	Capt	Maj	LtCol
Corpus Models	0.52	0.55	0.56	0.54	0.49
Pre-Corpus Models	0.52	0.54	0.55	0.53	0.48
Combined Score	0.55	0.57	0.58	0.56	0.50
Stacking elastic net	0.61	0.62	0.67	0.64	0.51

The ensemble greatly increases predictive power from 45% to 67%; however, it is not a practical technique. A FitRep reader only has a single glance of a word picture and compares the comments against prior experience or general perception. He or she does not have the ability to run a model of models of optimized term-document matrix configurations to interpret the quality of the comments relating to the MRO’s tier classification.

3. Power words

We find that the most powerful predictive model for performance tier classification is the elastic net with up to three word token. Through a series of penalties, the elastic net offers coefficients to words with the most predictive power. These power words are organized into the three categories: have no effect on predictability when the coefficient is driven to 0, “presence of” assists prediction with a positive coefficient, and “absence of” contributes to prediction with a negative coefficient. The magnitude of the coefficient speaks to how much it contributes to the model. We re-run the models for

each rank with optimal configurations to extract the power words. Tables 20 - 24 capture our top 20 positive and negative coefficients by rank per performance tier.

Table 20. Most Important Words through Presence and Absence for 2ndLt by Performance Tier Using Elastic Net

Top Tier				Middle Tier				Bottom Tier			
Presence of	**Coef**	**Absence of**	**Coef**	**Presence of**	**Coef**	**Absence of**	**Coef**	**Presence of**	**Coef**	**Absence of**	**Coef**
acumen	0.4817	orm	-0.2819	fight	0.5129	some	-0.4353	assimil	0.4273	mindset	-0.4251
simultan	0.4622	practic	-0.2796	grown	0.3695	fwd	-0.3790	orm	0.4188	cobra	-0.3620
prai	0.4253	zone	-0.2542	regardless	0.3498	charismat	-0.3704	refin	0.3398	marsoc	-0.3476
again	0.4208	fight	-0.2539	marsoc	0.3386	stellar	-0.2520	some	0.3259	agil	-0.3392
charismat	0.4100	reserv	-0.2516	agil	0.3006	human	-0.2489	join	0.2594	spectrum	-0.3105
among	0.3915	fex	-0.2376	univ	0.2973	assimil	-0.2472	fex	0.2146	logistician	-0.2987
spectrum	0.3667	between	-0.1946	battl	0.2687	prai	-0.2292	multitud	0.2130	regardless	-0.2951
thing	0.3607	submiss	-0.1819	between	0.2518	earliest	-0.2046	spent	0.1875	simultan	-0.2938
cobra	0.3240	assimil	-0.1801	zone	0.2494	acumen	-0.1950	learn	0.1803	univ	-0.2891
truli	0.3208	join	-0.1785	bias	0.2306	water	-0.1905	pcs	0.1766	acumen	-0.2867
abov	0.3126	alreadi	-0.1679	facet	0.2284	reinforc	-0.1829	meet	0.1644	except	-0.2764
except	0.3051	learn	-0.1599	defici	0.2139	trade	-0.1826	fwd	0.1597	compo	-0.2736
seri	0.3041	refin	-0.1596	collat	0.2060	keep	-0.1817	cycl	0.1456	mcas	-0.2733
mcas	0.3014	feedback	-0.1582	solver	0.2040	refin	-0.1803	stellar	0.1442	attain	-0.2723
top	0.2941	grown	-0.1464	practic	0.1989	counterpart	-0.1802	pickett	0.1433	among	-0.2718
highest	0.2921	young	-0.1458	submiss	0.1970	seri	-0.1753	note	0.1394	again	-0.2697
matur	0.2771	bias	-0.1438	mindset	0.1942	worth	-0.1729	test	0.1345	fight	-0.2590
water	0.2752	facet	-0.1431	incr	0.1904	thing	-0.1688	requisit	0.1334	highest	-0.2298
earliest	0.2573	squad	-0.1267	feedback	0.1856	simultan	-0.1684	question	0.1332	pressur	-0.2279
rise	0.2520	identifi	-0.1264	pressur	0.1538	wait	-0.1603	compress	0.1308	grown	-0.2231

Table 21. Most Important Words through Presence and Absence for 1stLt by Performance Tier Using Elastic Net

Top Tier				Middle Tier				Bottom Tier			
Presence of	**Coef**	**Absence of**	**Coef**	**Presence of**	**Coef**	**Absence**	**Coef**	**Presence of**	**Coef**	**Absence of**	**Coef**
top	0.3727	welcom	-0.3054	sent	0.3178	forth	-0.3186	seri	0.3333	must	-0.4693
finest	0.3706	seri	-0.2985	welcom	0.3107	rehear	-0.2689	quiet	0.2587	earliest	-0.4351
ahead	0.3545	sent	-0.2356	ran	0.2949	facet	-0.2495	adver	0.2500	top	-0.4198
must	0.3168	scout	-0.2215	highlight	0.2870	unparallel	-0.1988	forth	0.2426	highest	-0.3399
unmatch	0.3124	adjust	-0.2180	replac	0.2508	finest	-0.1814	learn	0.2118	name	-0.3209
earliest	0.2835	requisit	-0.2133	question	0.2380	prium	-0.1809	contemporar	0.2088	ahead	-0.2805
highest	0.2753	vigor	-0.1735	composur	0.2276	there	-0.1549	lack	0.2040	replac	-0.2656
name	0.2700	quiet	-0.1662	contributor	0.1977	reproach	-0.1444	facil	0.1852	addendum	-0.2452
facet	0.2657	trainer	-0.1660	heavi	0.1903	facil	-0.1443	grow	0.1848	unmatch	-0.2282
whom	0.2572	question	-0.1644	list	0.1889	network	-0.1415	promi	0.1796	student	-0.2253
now	0.2505	third	-0.1602	held	0.1826	just	-0.1311	watch	0.1756	enthusiast	-0.2158
cont	0.2375	held	-0.1599	reduc	0.1726	head	-0.1287	elig	0.1743	list	-0.2150
head	0.2357	updat	-0.1511	check	0.1724	profil	-0.1253	rehear	0.1665	expert	-0.2137
apart	0.2273	learn	-0.1470	student	0.1648	lack	-0.1228	updat	0.1661	cont	-0.2060
except	0.2236	rapid	-0.1469	contagi	0.1623	wealth	-0.1222	just	0.1598	fob	-0.1981
ive	0.2158	loyal	-0.1466	che	0.1533	adver	-0.1198	gener	0.1516	meticul	-0.1967
unparallel	0.2147	new	-0.1341	must	0.1525	fitrep	-0.1177	threat	0.1499	abov	-0.1965
sangin	0.1930	adver	-0.1302	requisit	0.1516	whom	-0.1136	awar	0.1471	except	-0.1957
page	0.1917	threat	-0.1278	earliest	0.1516	style	-0.1134	typic	0.1471	page	-0.1934
reproach	0.1903	myriad	-0.1277	copilot	0.1463	rai	-0.1121	shown	0.1388	finest	-0.1892

Table 22. Most Important Words through Presence and Absence for Capt by Performance Tier Using Elastic Net

Top Tier				Middle Tier				Bottom Tier			
Presence of	Coef	Absence of	Coef	Presence of	Coef	Absence of	Coef	Presence of	Coef	Absence of	Coef
top	0.4599	writer	-0.3299	fact	0.4388	breadth	-0.2506	adver	0.3475	top	-0.4174
cost	0.3881	progress	-0.2909	conus	0.2795	smooth	-0.2299	writer	0.3466	saw	-0.3504
number	0.3834	good	-0.2847	waiv	0.2669	shown	-0.2077	good	0.2542	ahead	-0.3438
ahead	0.3588	young	-0.2483	kcj	0.2513	occa	-0.2015	contemporari	0.2439	unflapp	-0.3239
must	0.3456	waiv	-0.2307	vigor	0.2097	relev	-0.1972	appar	0.2346	must	-0.3180
unflapp	0.3289	helicopt	-0.2188	trustworthi	0.1760	macg	-0.1734	wingman	0.2201	enthusiast	-0.2965
highest	0.3000	tbs	-0.2141	heavi	0.1742	delay	-0.1690	grow	0.1984	fact	-0.2956
finest	0.2976	appar	-0.2080	everyth	0.1537	fitrep	-0.1471	bold	0.1969	number	-0.2868
ever	0.2928	grow	-0.1782	saw	0.1517	nononsen	-0.1403	rapport	0.1763	everyth	-0.2687
absolut	0.2903	wingman	-0.1607	notch	0.1515	defen	-0.1367	progress	0.1749	highest	-0.2643
surpass	0.2649	contemporari	-0.1581	still	0.1491	caus	-0.1339	eager	0.1736	never	-0.2605
inspir	0.2618	incid	-0.1572	savvi	0.1474	asset	-0.1318	shown	0.1716	trustworthi	-0.2573
definit	0.2501	growth	-0.1567	curriculum	0.1466	levelhead	-0.1259	much	0.1590	expert	-0.2568
enthusiast	0.2464	valuabl	-0.1544	assur	0.1419	via	-0.1171	taught	0.1560	finest	-0.2484
embassi	0.2456	scope	-0.1538	main	0.1416	add	-0.1163	assimil	0.1511	surpass	-0.2424
tangibl	0.2348	abli	-0.1528	speak	0.1384	osd	-0.1163	learn	0.1365	electron	-0.2414
forget	0.2327	solid	-0.1465	cnaf	0.1369	opso	-0.1065	awar	0.1353	cnaf	-0.2359
commendatori	0.2215	fact	-0.1432	young	0.1333	find	-0.1025	genuin	0.1257	embassi	-0.2190
unwav	0.2183	promot	-0.1428	pictur	0.1304	draft	-0.1021	relief	0.1232	earliest	-0.2153
rare	0.2149	sole	-0.1425	fall	0.1299	top	-0.0993	abli	0.1176	commendatori	-0.2142

Table 23. Most Important Words through Presence and Absence for Maj by Performance Tier Using Elastic Net

Top Tier				Middle Tier				Bottom Tier			
Presence of	Coef	Absence of	Coef	Presence of	Coef	Absence of	Coef	Presence of	Coef	Absence of	Coef
tls	0.5069	conscienti	-0.4418	fulfil	0.2564	contemporari	-0.3588	popul	0.5078	tls	-0.4478
finest	0.3915	popul	-0.3761	elig	0.2475	hone	-0.3540	contemporari	0.4533	apart	-0.2847
apart	0.3644	elig	-0.2424	yet	0.2334	whom	-0.2664	conscienti	0.4250	phenomen	-0.2491
number	0.3519	warrior	-0.2112	attain	0.1784	adver	-0.2346	adver	0.4050	uncanni	-0.2314
phenomen	0.3021	formal	-0.1777	jun	0.1736	finest	-0.2042	hone	0.3208	top	-0.2299
ahead	0.2853	adver	-0.1703	spent	0.1629	quantico	-0.2038	post	0.2521	highest	-0.1913
highest	0.2416	solid	-0.1643	vmm	0.1617	repeat	-0.1918	against	0.2189	finest	-0.1873
uncanni	0.2367	fulfil	-0.1620	normal	0.1306	post	-0.1899	accord	0.1993	ahead	-0.1838
top	0.2333	count	-0.1540	credibl	0.1305	obtain	-0.1832	evaluat	0.1968	csc	-0.1825
repeat	0.2241	various	-0.1431	span	0.1262	crucial	-0.1826	age	0.1838	exceed	-0.1800
front	0.2149	attain	-0.1430	venu	0.1165	forum	-0.1822	ship	0.1688	stellar	-0.1748
innat	0.2078	qualif	-0.1398	warrior	0.1136	number	-0.1783	good	0.1680	number	-0.1736
obtain	0.2028	against	-0.1376	csc	0.1096	vma	-0.1484	oversaw	0.1674	enthusiast	-0.1710
exceed	0.1945	jun	-0.1367	polish	0.1071	ship	-0.1482	rbe	0.1651	unmatch	-0.1686
ive	0.1910	accord	-0.1320	attent	0.1036	evaluat	-0.1340	part	0.1629	must	-0.1672
forum	0.1853	rbe	-0.1279	readili	0.1033	kcj	-0.1334	smcr	0.1604	now	-0.1613
exact	0.1762	did	-0.1279	notic	0.1018	popul	-0.1317	qualif	0.1572	keen	-0.1566
now	0.1675	adroit	-0.1268	just	0.1013	add	-0.1307	upcom	0.1563	break	-0.1542
push	0.1665	retent	-0.1238	innovat	0.1009	exact	-0.1291	quantico	0.1523	front	-0.1443
stellar	0.1659	cando	-0.1158	count	0.1001	welcom	-0.1230	solid	0.1451	innat	-0.1441

Table 24. Most Important Words through Presence and Absence for LtCol by Performance Tier Using Elastic Net

Top Tier				Middle Tier				Bottom Tier			
Presence of	Coef	Absence of	Coef	Presence of	Coef	Absence of	Coef	Presence of	Coef	Absence of	Coef
top	0.2969	cprl	-0.5157	cprl	0.8268	breadth	-0.2506	low	0.3684	cprl	-0.3111
add	0.2562	left	-0.3115	left	0.4553	smooth	-0.2299	breadth	0.3594	enthusiast	-0.2085
find	0.2429	low	-0.2750	handson	0.2530	shown	-0.2077	caus	0.3332	top	-0.1976
devot	0.2305	palm	-0.2517	almost	0.2426	occa	-0.2015	osd	0.2665	import	-0.1968
dramat	0.1931	satisfi	-0.2178	flag	0.2258	relev	-0.1972	retain	0.2497	almost	-0.1467
via	0.1866	flag	-0.2034	origin	0.2135	macg	-0.1734	shown	0.2495	devot	-0.1443
import	0.1856	caus	-0.1993	center	0.1956	delay	-0.1690	opso	0.2459	left	-0.1437
macg	0.1834	injur	-0.1939	satisfi	0.1843	fitrep	-0.1471	injur	0.2330	find	-0.1404
enthusiast	0.1704	che	-0.1897	inspector	0.1764	nononsen	-0.1403	che	0.2020	add	-0.1399
there	0.1583	streamlin	-0.1857	refin	0.1741	defen	-0.1367	fitrep	0.1825	handson	-0.1352
defen	0.1521	multipli	-0.1789	palm	0.1699	caus	-0.1339	smooth	0.1825	origin	-0.1317
etc	0.1205	solid	-0.1700	remark	0.1572	asset	-0.1318	final	0.1747	readi	-0.1256
plus	0.1161	subsequ	-0.1607	lce	0.1561	levelhead	-0.1259	help	0.1692	remark	-0.1183
delay	0.1112	retain	-0.1570	post	0.1426	via	-0.1171	multipli	0.1649	dramat	-0.0985
platoon	0.1081	osd	-0.1503	tireless	0.1407	add	-0.1163	relev	0.1611	col	-0.0966
along	0.1025	due	-0.1422	trainer	0.1382	osd	-0.1163	dozen	0.1561	whole	-0.0915
abli	0.0949	center	-0.1405	whole	0.1260	opso	-0.1065	solid	0.1554	trainer	-0.0881
sourc	0.0882	opso	-0.1394	builder	0.1133	find	-0.1025	occa	0.1419	tireless	-0.0832
asset	0.0881	inspector	-0.1257	optim	0.1115	draft	-0.1021	due	0.1346	colonel	-0.0805
colonel	0.0870	dozen	-0.1255	welcom	0.1089	top	-0.0993	congress	0.1272	refin	-0.0722

V. CONCLUSION AND RECOMMENDATIONS

The motivation of the study is to determine whether the text fields provide additional predictive information on the performance of a MRO. The old PES manual provided structure to the comments because they were more valuable than the marks assignments. The new PES manual puts less direction on the text comments by stating that they should provide a "more complete and detailed evaluation of the MRO's professional character and may address any entry made in sections A through H or as the Reporting Senior deems appropriate" (Commandant of the Marine Corps, 2015, p. 4-39). Furthermore, the PES manual seeks to ensure that the comments are consistent with the marks. For these reasons, we expect to find consistency between like-tiered individuals in the comment boxes. Although a redundant field would not benefit predictive power, the use of words to clarify, enhance, or complement the markings would facilitate performance classification, promotion, and future assignment.

We start with the analysis of the distribution of the comparative assessments and relative values to assess the quality of our response variable. The latter show that while there is an attempt to balance fitness reports across a scale from 80 to 100, the distributions are skewed and concentrated in a narrow range. As a Marine progresses from 2ndLt to LtCol, the marks go from a right skew to a left skew. These asymmetries are caused by a few outliers, either outstanding or terrible performers, and allow the RS to mask concentrations of fitness reports in a small range. Additionally, the ROs do not follow the "Christmas tree" distribution prescribed in the PES Manual. For each rank, no less than 70% of MROs were contained in two of eight blocks. These concentrations of RS and RO markings make mathematical breakout of Marines more difficult and place higher value on the contents of the textual fields.

Similarly, by assessing concurrence between the RS and RO evaluations, we find that although ROs indicate formal non-concurrence with the RSs' evaluations 0.01% of the time, their tier assessments disagree 49% of the time. Although the scope of the RO's evaluation is to examine the MRO's performance and potential as whole, by concurring with the report, he or she agrees that the marks are appropriate to the MRO's

performance and are not inflated. While some overlap between the tier borders is expected, it is not expected up to 51% of the time.

By investigating directed comments, we find that they are not consistently used in accordance with the PES manual. An important factor on whether a comment is made, is whether a prompt is given when writing a particular section of the fitness report. When a prompt is not given, the rate of providing comments is low as seen in aviator's flying proficiency (2.37%), an MRO's compliance with ORM guidance (2.34%), and fitness reports less than 90 days (13.4%). When a prompt is given, comments are provided more frequently, such as duty in a combat zone (62.1%) and awards (81.2%). Additionally, the RO often allows the administratively incorrect fitness report to be forwarded to Headquarters Marine Corps. While the resources to make appropriate comments are available, training and education to address the proper use of directed comments are lacking.

Using prior work on Amazon product reviews and professor evaluations to guide our pre-corpus analysis of the FitRep comments, we derive a set of metrics of writing quality utilizing spelling errors, five readability statistics, word tallies, and character counts. With those, we correctly classify 48% of fitness reports with respect to tier groups, which is 15% better than naïve assignment. We find that while the rates of spelling errors, Flesch Kincaid Index, Flesch Grade Level, SMOG Index, and the Colman Liau Index are statistically significant predictors, they are not practically significant. We regard a predictor variable as having practical significance if it appears consistently in predictive modeling. Our very large sample size gives statistical tests high power for detecting predictive ability. The word counts and the ARI index are statistically and practically significant and provide the best predictive power. Simple analysis of the descriptive aspect of the comments reveals that simple words woven into well written, long sentences with few spelling mistakes are indicators of the MRO belonging to a higher performing tier. Conversely, long words in complicated sentences with many spelling errors are key indicators of bottom-tier MRO.

We examine the structure of the comments and the use of power words. Although the PES manual does not suggest comment block structure beyond mandatory, directed,

and additional comments, the reporting chain has tended to prefer the old PES manual style of providing opening remarks, comments on performance and character, and closing with comments on promotion, retention, and assignment. This conformance to the old manual could be due to the RSs and ROs who operated on the old system filling the gap of education and specific guidance in the PES. The matching also coincides with the lack of guidance on how to write fitness reports. Nearly every Marine Corps installation bookstore sells the “Fitness Report Writing Guide for Marines” (Drewry, 1998) , but it was originally written in 1986 and has not been republished since 1998.

We search the corpus of fitness reports for meaningful correlation between the words “promotion,” “retention,” “command,” “potential,” and words indicating various types of assignments. These words are important because the Marine Corps directs evaluators to use them in the comment fields of fitness reports. We find that adjectives that tend to be strongly correlated with these words for one tier tend to be similarly correlated across all tiers. For example, the term “unlimited growth potential” occurs with the highest frequency in each of the three tiers. We find that positive comments on future command assignments are reserved for top tier Marines and that the word “peer” is mostly earmarked for the bottom tier. The lack of correlation between the keywords and qualifying adjectives render classification of performance by a interested reader without statistical machine learning techniques difficult. Although one may expect that intra-text rankings such as “#1 Officer” or “Best Capt” to be strong indicators of the highest performers, the phrase is formally accurate only 35% of the time for RS evaluations. When an MRO receives a comment such as “number 1 Marine” from an RO, he or she is with 61% of all officers in the same grade. These observations are due in part to the concentration of evaluation marks along a narrow performance range.

We use seven supervised machine-learning algorithms on 360 term-document matrices to find the model configuration with the most predictive power. Using the harmonic mean between recall and precision, the best minimum-tier correct classification rate is 45%; however, by using a penalty-enhanced generalized linear model of the seven best model predictions, we increase our correct classification rate to 67%. The best term-document matrix configuration is for traditionally weighted matrices of single-word

tokens. The use of single-word tokens demonstrates that, contrary to literature on predictive text mining such as Chuang et al. (2012), the interactions between key terms (e.g. “must promote” or “highest recommendation,” etcetera) were not useful in tier prediction. Using the best performing predictor, which is a penalty enhanced generalized linear model, we extract the “power words” and their weights to determine the most useful predictive terms. The terms represented in Tables 20 – 24 exhibit a pattern of performance-related superlatives. Words such as “highest,” “unmatched,” “top,” and “enthusiastic” are reserved for the top tier while “contemporaries,” “peers,” and “qualify” are associated with the bottom tier.

To provide more value to the textual information contained in FitReps requires examining the marks and comments. First, the relative values and comparative assessments should conform more closely to the distributions indicated in the PES manual. An alternative method is to implement an intra-grade ranking. By ranking everyone from 1 (highest) to *n* (lowest) enables the Marine Corps to set cutoffs for promotion and assignment boards.

The second recommendation is for the Marine Corps to publish specific guidance, by rank and tier, of what words the Marine Corps deems appropriate for evaluation purposes. This guidance should be reinforced by an accountability system to ensure RSs and ROs do not use words reserved for a tier group other than the intended one. The Marine Corps should endorse the publication of an updated version of “Fitness Report Writing Guide for Marines” to synchronize FitRep writers on a common performance vocabulary.

To check the administrative compliance of the fitness reports, Manpower and Reserve Affair’s Records and Performance Branch should conduct a more thorough administrative review process to ensure compliance with Marine Corps guidance, policy, and direction. Incorrect fitness reports should be returned to the reviewing officer for correction since he or she is the one who is supposed to certify the administrative correctness of the fitness reports. This review process will have the additional indirect benefit of providing training for the reviewing officer.

Finally, as recommended by Clemens et al. (2012), the Marine Corps should increase its investment in the officers' fitness report writing training. As the primary tool for evaluating the retention, promotion, and assignment of Marines, the fitness report should be interwoven with an officer's continuing education such as through Marinenet or incorporated into each grade's PME advancement. The Marine Corps should reinforce distributions of assigned marks and address the duties and responsibilities of officers who assume RO responsibilities.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. FITNESS REPORT EXAMPLE

USMC FITNESS REPORT (1610)
NAVMC 10835A (Rev. 1-01)(P
PREVIOUS EDITIONS WILL NOT BE USED
FOUO - Privacy sensitive if filled in

DO NOT STAPLE THIS FORM

COMMANDANT'S GUIDANCE

The completed fitness report is the most important information component in manpower management. It is the primary means of evaluating a Marine's performance and is the Commandant's primary tool for the selection of personnel for promotion, augmentation, resident schooling, command, and duty assignments. Therefore, the completion of this report is one of an officer's most critical responsibilities. Inherent in this duty is the commitment of each Reporting Senior and Reviewing Officer to ensure the integrity of the system by giving close attention to accurate marking and timely reporting. Every officer serves a role in the scrupulous maintenance of this evaluation system, ultimately important to both the individual and the Marine Corps. Inflationary markings only serve to dilute the actual value of each report. Reviewing Officers will not concur with inflated reports.

A. ADMINISTRATIVE INFORMATION

1. Marine Reported On:

a. Last Name	b. First Name	c. MI	d. ID	e. Grade	f. DOR	g. PMOS	h. BILMOS

2. Organization:

a. MCC	b. RUC	c. Unit Description

3. Occasion and Period Covered:

a. OCC	b. From	To	c. Type

4. Duty Assignment (descriptive title):

5. Special Case: a. Adverse ☐ b. Not Observed ☐ c. Extended ☐

6. Marine Subject Of: a. Commendatory Material ☐ b. Derogatory Material ☐ c. Disciplinary Action ☐

7. Recommended For Promotion: a. Yes ☐ b. No ☐ c. N/A ☐

8. Special Information:

a. QUAL		d. HT(in.)		g. Reserve Component	
b. PFT		e. WT		h. Status	
c. CFT		f. Body Fat		i. Future Use	

9. Duty Preference:

	a. Code	b. Descriptive Title
1st		
2nd		
3rd		

10. Reporting Senior:

a. Last Name	b. Init	c. Service	d. ID	e. Grade	f. Duty Assignment

11. Reviewing Officer:

a. Last Name	b. Init	c. Service	d. ID	e. Grade	f. Duty Assignment

B. BILLET DESCRIPTION

C. BILLET ACCOMPLISHMENTS

1. Marine Reported On: a. Last Name	b. First Name	c. MI	d. ID	2. Occasion and Period Covered: a. OCC	b. From	To

D. MISSION ACCOMPLISHMENT

1. PERFORMANCE. Results achieved during the reporting period. How well those duties inherent to a Marine's billet, plus all additional duties, formally and informally assigned, were carried out. Reflects a Marine's aptitude, competence, and commitment to the unit's success above personal reward. Indicators are time and resource management, task prioritization, and tenacity to achieve positive ends consistently.

ADV							N/O
	Meets requirements of billet and additional duties. Aptitude, commitment, and competence meet expectations. Results maintain status quo.		Consistently produces quality results while measurably improving unit performance. Habitually makes effective use of time and resources; improves billet procedures and products. Positive impact extends beyond billet expectations.		Results far surpass expectations. Recognizes and exploits new resources; creates opportunities. Emulated; sought after as an expert with influence beyond unit. Impact significant; innovative approaches to problems produce significant gains in quality and efficiency.		
A	B	C	D	E	F	G	H
☐	☐	☐	☐	☐	☐	☐	☐

2. PROFICIENCY. Demonstrates technical knowledge and practical skill in the execution of the Marine's overall duties. Combines training, education and experience. Translates skills into actions which contribute to accomplishing tasks and missions. Imparts knowledge to others. Grade dependent.

ADV							N/O
	Competent. Possesses the requisite range of skills and knowledge commensurate with grade and experience. Understands and articulates basic functions related to mission accomplishment.		Demonstrates mastery of all required skills. Expertise, education and experience consistently enhance mission accomplishment. Innovative troubleshooter and problem solver. Effectively imparts skills to subordinates.		True expert in field. Knowledge and skills impact far beyond those of peers. Translates broad-based education and experience into forward thinking, innovative actions. Makes immeasurable impact on mission accomplishment. Peerless teacher, selflessly imparts expertise to subordinates, peers, and seniors.		
A	B	C	D	E	F	G	H
☐	☐	☐	☐	☐	☐	☐	☐

JUSTIFICATION:

E. INDIVIDUAL CHARACTER

1. COURAGE. Moral or physical strength to overcome danger, fear, difficulty or anxiety. Personal acceptance of responsibility and accountability, placing conscience over competing interests regardless of consequences. Conscious, overriding decision to risk bodily harm or death to accomplish the mission or save others. The will to persevere despite uncertainty.

ADV							N/O
	Demonstrates inner strength and acceptance of responsibility commensurate with scope of duties and experience. Willing to face moral or physical challenges in pursuit of mission accomplishment.		Guided by conscience in all actions. Proven ability to overcome danger, fear, difficulty or anxiety. Exhibits bravery in the face of adversity and uncertainty. Not deterred by morally difficult situations or hazardous responsibilities.		Uncommon bravery and capacity to overcome obstacles and inspire others in the face of moral dilemma or life-threatening danger. Demonstrated under the most adverse conditions. Selfless. Always places conscience over competing interests regardless of physical or personal consequences.		
A	B	C	D	E	F	G	H
☐	☐	☐	☐	☐	☐	☐	☐

2. EFFECTIVENESS UNDER STRESS. Thinking, functioning and leading effectively under conditions of physical and/or mental pressure. Maintaining composure appropriate for the situation, while displaying steady purpose of action, enabling one to inspire others while continuing to lead under adverse conditions. Physical and emotional strength, resilience and endurance are elements.

ADV							N/O
	Exhibits discipline and stability under pressure. Judgment and effective problem-solving skills are evident.		Consistently demonstrates maturity, mental agility and willpower during periods of adversity. Provides order to chaos through the application of intuition, problem-solving skills, and leadership. Composure reassures others.		Demonstrates seldom-matched presence of mind under the most demanding circumstances. Stabilizes any situation through the resolute and timely application of direction, focus and personal presence.		
A	B	C	D	E	F	G	H
☐	☐	☐	☐	☐	☐	☐	☐

3. INITIATIVE. Action in the absence of specific direction. Seeing what needs to be done and acting without prompting. The instinct to begin a task and follow through energetically on one's own accord. Being creative, proactive and decisive. Transforming opportunity into action.

ADV							N/O
	Demonstrates willingness to take action in the absence of specific direction. Acts commensurate with grade, training and experience.		Self-motivated and action-oriented. Foresight and energy consistently transform opportunity into action. Develops and pursues creative, innovative solutions. Acts without prompting. Self-starter.		Highly motivated and proactive. Displays exceptional awareness of surroundings and environment. Uncanny ability to anticipate mission requirements and quickly formulate original, far-reaching solutions. Always takes decisive, effective action.		
A	B	C	D	E	F	G	H
☐	☐	☐	☐	☐	☐	☐	☐

JUSTIFICATION:

NAVMC 10835B (Rev. 4-03) (P

1. Marine Reported On: a. Last Name	b. First Name	c. MI	d. ID	2. Occasion and Period Covered: a. OCC	b. From	To

F. LEADERSHIP

1. LEADING SUBORDINATES. The inseparable relationship between leader and led. The application of leadership principles to provide direction and motivate subordinates. Using authority, persuasion and personality to influence subordinates to accomplish assigned tasks. Sustaining motivation and morale while maximizing subordinates' performance.

ADV	B	D	F	N/O
	Engaged; provides instructions and directs execution. Seeks to accomplish mission in ways that sustain motivation and morale. Actions contribute to unit effectiveness.	Achieves a highly effective balance between direction and delegation. Effectively tasks subordinates and clearly delineates standards expected. Enhances performance through constructive supervision. Fosters motivation and enhances morale. Builds and sustains teams that successfully meet mission requirements. Encourages initiative and candor among subordinates.	Promotes creativity and energy among subordinates by striking the ideal balance of direction and delegation. Achieves highest levels of performance from subordinates by encouraging individual initiative. Engenders willing subordination, loyalty, and trust that allow subordinates to overcome their perceived limitations. Personal leadership fosters highest levels of motivation and morale, ensuring mission accomplishment even in the most difficult circumstances.	

A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐

2. DEVELOPING SUBORDINATES. Commitment to train, educate, and challenge all Marines regardless of race, religion, ethnic background, or gender. Mentorship. Cultivating professional and personal development of subordinates. Developing team players and esprit de corps. Ability to combine teaching and coaching. Creating an atmosphere tolerant of mistakes in the course of learning.

ADV	B	D	F	N/O
	Maintains an environment that allows personal and professional development. Ensures subordinates participate in all mandated development programs.	Develops and institutes innovative programs, to include PME, that emphasize personal and professional development of subordinates. Challenges subordinates to exceed their perceived potential thereby enhancing unit morale and effectiveness. Creates an environment where all Marines are confident to learn through trial and error. As a mentor, prepares subordinates for increased responsibilities and duties.	Widely recognized and emulated as a teacher, coach and leader. Any Marine would desire to serve with this Marine because they know they will grow personally and professionally. Subordinate and unit performance far surpassed expected results due to MRO's mentorship and team building talents. Attitude toward subordinate development is infectious, extending beyond the unit.	

A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐

3. SETTING THE EXAMPLE. The most visible facet of leadership: how well a Marine serves as a role model for all others. Personal action demonstrates the highest standards of conduct, ethical behavior, fitness, and appearance. Bearing, demeanor, and self-discipline are elements.

ADV	B	D	F	N/O
	Maintains Marine Corps standards for appearance, weight, and uniform wear. Sustains required level of physical fitness. Adheres to the tenets of the Marine Corps core values.	Personal conduct on and off duty reflects highest Marine Corps standards of integrity, bearing and appearance. Character is exceptional. Actively seeks self-improvement in wide-ranging areas. Dedication to duty and professional example encourage others' self-improvement efforts.	Model Marine, frequently emulated. Exemplary conduct, behavior, and actions are tone-setting. An inspiration to subordinates, peers, and seniors. Remarkable dedication to improving self and others.	

A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐

4. ENSURING WELL-BEING OF SUBORDINATES. Genuine interest in the well-being of Marines. Efforts enhance subordinates' ability to concentrate/focus on unit mission accomplishment. Concern for family readiness is inherent. The importance placed on welfare of subordinates is based on the belief that Marines take care of their own.

ADV	B	D	F	N/O
	Deals confidently with issues pertinent to subordinate welfare and recognizes suitable courses of action that support subordinates' well-being. Applies available resources, allowing subordinates to effectively concentrate on the mission.	Instills and/or reinforces a sense of responsibility among junior Marines for themselves and their subordinates. Actively fosters the development of and uses support systems for subordinates which improve their ability to contribute to unit mission accomplishment. Efforts to enhance subordinate welfare improve the unit's ability to accomplish its mission.	Noticeably enhances subordinates well-being, resulting in a measurable increase in unit effectiveness. Maximizes unit and base resources to provide subordinates with the best support available. Proactive approach serves to energize unit members to "take care of their own," thereby correcting potential problems before they can hinder subordinates' effectiveness. Widely recognized for techniques and policies that produce results and build morale. Builds strong family atmosphere. Puts motto *Mission first, Marines always*, into action.	

A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐

5. COMMUNICATION SKILLS. The efficient transmission and receipt of thoughts and ideas that enable and enhance leadership. Equal importance given to listening, speaking, writing, and critical reading skills. Interactive, allowing one to perceive problems and situations, provide concise guidance, and express complex ideas in a form easily understood by everyone. Allows subordinates to ask questions, raise issues and concerns and venture opinions. Contributes to a leader's ability to motivate as well as counsel.

ADV	B	D	F	N/O
	Skilled in receiving and conveying information. Communicates effectively in performance of duties.	Clearly articulates thoughts and ideas, verbally and in writing. Communication in all forms is accurate, intelligent, concise, and timely. Communicates with clarity and verve, ensuring understanding of intent or purpose. Encourages and considers the contributions of others.	Highly developed facility in verbal communication. Adept in composing written documents of the highest quality. Combines presence and verbal skills which engender confidence and achieve understanding irrespective of the setting, situation, or size of the group addressed. Displays an intuitive sense of when and how to listen.	

A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐

JUSTIFICATION:

1. Marine Reported On: a. Last Name	b. First Name	c. MI	d. ID	2. Occasion and Period Covered: a. OCC	b. From	To

G. INTELLECT AND WISDOM

1.PROFESSIONAL MILITARY EDUCATION (PME). Commitment to intellectual growth in ways beneficial to the Marine Corps. Increases the breadth and depth of warfighting and leadership aptitude. Resources include resident schools; professional qualifications and certification processes; nonresident and other extension courses; civilian educational institution coursework; a personal reading program that includes (but is not limited to) selections from the Commandant's Reading List; participation in discussion groups and military societies; and involvement in learning through new technologies.

ADV							N/O
	Maintains currency in required military skills and related developments. Has completed or is enrolled in appropriate level of PME for grade and level of experience. Recognizes and understands new and creative approaches to service issues. Remains abreast of contemporary concepts and issues.		PME outlook extends beyond MOS and required education. Develops and follows a comprehensive personal program which includes broadened professional reading and/or academic course work; advances new concepts and ideas.		Dedicated to life-long learning. As a result of active and continuous efforts, widely recognized as an intellectual leader in professionally related topics. Makes time for study and takes advantage of all resources and programs. Introduces new and creative approaches to services issues. Engages in a broad spectrum of forums and dialogues.		
A ☐	B ☐	C ☐	D ☐	E ☐	F ☐	G ☐	H ☐

2. DECISION MAKING ABILITY. Viable and timely problem solution. Contributing elements are judgment and decisiveness. Decisions reflect the balance between an optimal solution and a satisfactory, workable solution that generates tempo. Decisions are made within the context of the commander's established intent and the goal of mission accomplishment. Anticipation, mental agility, intuition, and success are inherent.

ADV							N/O
	Makes sound decisions leading to mission accomplishment. Actively collects and evaluates information and weighs alternatives to achieve timely results. Confidently approaches problems; accepts responsibility for outcomes.		Demonstrates mental agility; effectively prioritizes and solves multiple complex problems. Analytical abilities enhanced by experience, education, and intuition. Anticipates problems and implements viable, long-term solutions. Steadfast, willing to make difficult decisions.		Widely recognized and sought after to resolve the most critical, complex problems. Seldom matched analytical and intuitive abilities; accurately foresees unexpected problems and arrives at well-timed decisions despite fog and friction. Completely confident approach to all problems. Masterfully strikes a balance between the desire for perfect knowledge and greater tempo.		
A ☐	B ☐	C ☐	D ☐	E ☐	F ☐	G ☐	H ☐

3. JUDGMENT. The discretionary aspect of decision making. Draws on core values, knowledge, and personal experience to make wise choices. Comprehends the consequences of contemplated courses of action.

ADV							N/O
	Majority of judgments are measured, circumspect, relevant and correct.		Decisions are consistent and uniformly correct, tempered by consideration of their consequences. Able to identify, isolate and assess relevant factors in the decision making process. Opinions sought by others. Subordinates personal interest in favor of impartiality.		Decisions reflect exceptional insight and wisdom beyond this Marine's experience. Counsel sought by all; often an arbiter. Consistent, superior judgment inspires the confidence of seniors.		
A ☐	B ☐	C ☐	D ☐	E ☐	F ☐	G ☐	H ☐

JUSTIFICATION:

H. FULFILLMENT OF EVALUATION RESPONSIBILITIES

1. EVALUATIONS. The extent to which this officer serving as a reporting official conducted, or required others to conduct, accurate, uninflated, and timely evaluations.

ADV							N/O
	Occasionally submitted untimely or administratively incorrect evaluations. As RS, submitted one or more reports that contained inflated markings. As RO, concurred with one or more reports from subordinates that were returned by HQMC for inflated marking.		Prepared uninflated evaluations which were consistently submitted on time. Evaluations accurately described performance and character. Evaluations contained no inflated markings. No reports returned by RO or HQMC for inflated marking. No subordinates' reports returned by HQMC for inflated marking. Few, if any, reports were returned by RO or HQMC for administrative errors. Section Cs were void of superlatives. Justifications were specific, verifiable, substantive, and where possible, quantifiable and supported the markings given.		No reports submitted late. No reports returned by either RO or HQMC for administrative correction or inflated markings. No subordinates' reports returned by HQMC for administrative correction or inflated markings. Returned procedurally or administratively incorrect reports to subordinates for correction. As RO nonconcurred with all inflated reports.		
A ☐	B ☐	C ☐	D ☐	E ☐	F ☐	G ☐	H ☐

JUSTIFICATION:

1. Marine Reported On: a. Last Name	b. First Name	c. MI	d. ID	2. Occasion and Period Covered: a. OCC	b. From	To

I. DIRECTED AND ADDITIONAL COMMENTS

J. CERTIFICATION

1. I CERTIFY that to the best of my knowledge and belief all entries made hereon are true and without prejudice or partiality and that I have provided a signed copy of this report to the Marine Reported on.

(Signature of Reporting Senior) (Date in YYYYMMDD format)

2. I ACKNOWLEDGE the adverse nature of this report and

☐ I have no statement to make

☐ I have attached a statement

(Signature of Marine Reported On) (Date in YYYYMMDD format)

K. REVIEWING OFFICER COMMENTS

1. OBSERVATION: ☐ Sufficient ☐ Insufficient

2. EVALUATION: ☐ Concur ☐ Do Not Concur

3. COMPARATIVE ASSESSMENT: Provide a comparative assessment of potential by placing an "X" in the appropriate box. In marking the comparison, consider all Marines of this grade whose professional abilities are known to you personally.

DESCRIPTION		COMPARATIVE ASSESSMENT
THE EMINENTLY QUALIFIED MARINE	☐	
ONE OF THE FEW EXCEPTIONALLY QUALIFIED MARINES	☐ ☐	
ONE OF THE MANY HIGHLY QUALIFIED PROFESSIONALS WHO FORM THE MAJORITY OF THIS GRADE	☐ ☐ ☐	
A QUALIFIED MARINE	☐	
UNSATISFACTORY	☐	

4. REVIEWING OFFICER COMMENTS: Amplify your comparative assessment mark; evaluate potential for continued professional development to include: promotion, command, assignment, resident PME, and retention; and put Reporting Senior marks and comments in perspective.

5. I CERTIFY that to the best of my knowledge and belief all entries made hereon are true and without prejudice or partiality.

(Signature of Reviewing Officer) (Date in YYYYMMDD format)

6. I ACKNOWLEDGE the adverse nature of this report and

☐ I have no statement to make

☐ I have attached a statement

(Signature of Marine Reported On) (Date in YYYYMMDD format)

L. ADDENDUM PAGE

ADDENDUM PAGE ATTACHED: ☐ YES

NAVMC 10835E (Rev. 4-03) (P A-PES 5.4.5.0) FOR OFFICIAL USE ONLY - Privacy sensitive when filled in. PAGE 5 OF 5

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. DATA FIELD DESCRIPTIONS

Table 25. The 28 Administrative Fields Contained in the FitRep Data Set

Field	Type of Data	# of levels	Categorical Variable Level
Fiscal Year	Categorical Variable	11	FY2006-FY2016
Case Number	Categorical Variable	4,761	3-5596
Received Date	Numeric		
Occasion	Categorical Variable	13	AN, AR, CD, CH, CS, DC, EN, FD, GC, RT, SA, TD, TR
Promotion Date	Numeric		
Grade	Categorical Variable	5	2ndLt, 1stLt, Capt, Maj, LtCol
From Date	Numeric		
To Date	Numeric		
Occasion	Categorical Variable		
Primary MOS	Categorical Variable	991	0111-9999 (Commandant of the Marine Corps, 2013)
Billet MOS	Categorical Variable	991	0111-9999 (Commandant of the Marine Corps, 2013)
Duty Type	Categorical Variable	3	Normal, Combat, Academic
Months	Numeric		3-13 (min observation time is 90 days at a minimum, FIREPS are annual but if an occasion of higher priority occurs within 30 days, the end date can be 13 months)
Marine Subject	Categorical Variable	2	Yes and No
Special Case	Categorical Variable	3	Observed, Not-Observed, Extended
Adverse	Categorical Variable	2	Yes and No
Duty Assignment	Character		
Monitor Control Code (unit assigned)	Categorical Variable	2483	3-digit alpha numeric combination
Reporting Unit Code (unit assigned)	Categorical Variable	2322	5-digit numeric value
Unit Description	Character		
Rifle and Pistol Qualification	Categorical Variable	25	Not required, Required-did-not-shoot, Unqualified, Marksman, Sharpshooter, Expert x 2
Physical Fitness Test	Numeric		0-300
Combat Fitness Test	Numeric		0-300
Height (inches)	Numeric		
Weight (pounds)	Numeric		
Body Fat (percent)	Numeric		

Table 26. MRO Performance Fields Contained in the FitRep Data Set

Section	Sub-Section	Type of Data	Categorical Variable Level
Mission	Performance	Categorical Variable	A-H
Mission	Proficiency	Categorical Variable	A-H
Individual	Courage	Categorical Variable	A-H
Individual	Character	Categorical Variable	A-H
Individual	Effectiveness under stress	Categorical Variable	A-H
Leadership	Leading Marines	Categorical Variable	A-H
Leadership	Developing Subordinates	Categorical Variable	A-H
Leadership	Setting the Example	Categorical Variable	A-H
Leadership	Ensure Well Being of Subordinates	Categorical Variable	A-H
Leadership	Communication Skills	Categorical Variable	A-H
Intellectual	Professional Military Education	Categorical Variable	A-H
Intellectual	Decision Making	Categorical Variable	A-H
Intellectual	Judgement	Categorical Variable	A-H
Evaluation	Fulfillment of evaluation	Categorical Variable	A-H

Table 27. Reporting Senior and Reviewing Officer Markings Contained in the FitRep Data Set

Field	Data Type	# of Levels	Categorical Variable Level / Numeric Range
RS Service	Categorical Variable	5	USMC, USN, USA, USAF, CIV
RS Grade	Categorical Variable	10	2ndLt- Gen
Recommended for Promotion	Categorical Variable	2	Yes / No
Order of Report	Numeric		Unrestrained non-negative integer
Total Number of Reports written	Numeric		Unrestrained non-negative integer
Report Raw Score Average	Numeric		0-5 Rational Number
RS Average	Numeric		0-5 Rational Number
RS Highest Score	Numeric		0-5 Rational Number
RS How many reports at high score	Numeric		0-5 Rational Number
Relative Value at Processing	Numeric		80-100 Rational Number
Relative Value, Cumulative	Numeric		80-100 Rational Number
RO Service	Categorical Variable	5	USMC, USN, USA, USAF, CIV
RO Grade	Categorical Variable	10	2ndLt - Gen
MRO Score assigned by RO	Categorical Variable	8	
Sufficient Observation	Categorical Variable	2	Sufficient / Insufficient
Concurrence with RO	Categorical Variable	2	Yes / No
# FitReps at RO Score 1 at Processing	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 2 at Processing	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 3 at Processing	Numeric		Unrestrained non-negative

Field	Data Type	# of Levels	Categorical Variable Level / Numeric Range
			integer
# FitReps at RO Score 4 at Processing	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 5 at Processing	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 6 at Processing	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 7 at Processing	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 8 at Processing	Numeric		Unrestrained non-negative integer
At Processing: Sum of count of observed scores	Numeric	$\sum ROMarkings$	
At Processing: # of FitReps with score x times score value	Numeric	$\sum ROMarkings * MROScore$	
At Processing: RO Report Average	Numeric	$\frac{\sum ROMarkings * MROScore}{\sum ROMarkings}$	
At Processing: Difference between RO average and Score	Numeric	ROVDiff	
# FitReps at RO Score 1, Cumulative	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 2, Cumulative	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 3, Cumulative	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 4, Cumulative	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 5, Cumulative	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 6, Cumulative	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 7, Cumulative	Numeric		Unrestrained non-negative integer
# FitReps at RO Score 8, Cumulative	Numeric		Unrestrained non-negative integer
Cumulative: Sum of count of observed scores	Numeric	$\sum ROMarkings$	
Cumulative: # of FitReps with score x times score value	Numeric	$\sum ROMarkings * MROScore$	
Cumulative: RO Report Average	Numeric	$\frac{\sum ROMarkings * MROScore}{\sum ROMarkings}$	
Cumulative: Difference between RO average and Score	Numeric	ROVDiff	

Table 28. Text Variables Contained in the FitRep Data Set

Section	Sub-Section	Type of Data
Section I	Case ID	Categorical Variable
	Date	Numeric
	Section I comments	Character
Section K	Case ID	Categorical Variable
	Date	Numeric
	Section K comments	Character
Addendum	Case ID	Categorical Variable
	Date	Numeric
	Addendum	Character

APPENDIX C. UNSUPERVISED WORD CORRELATION

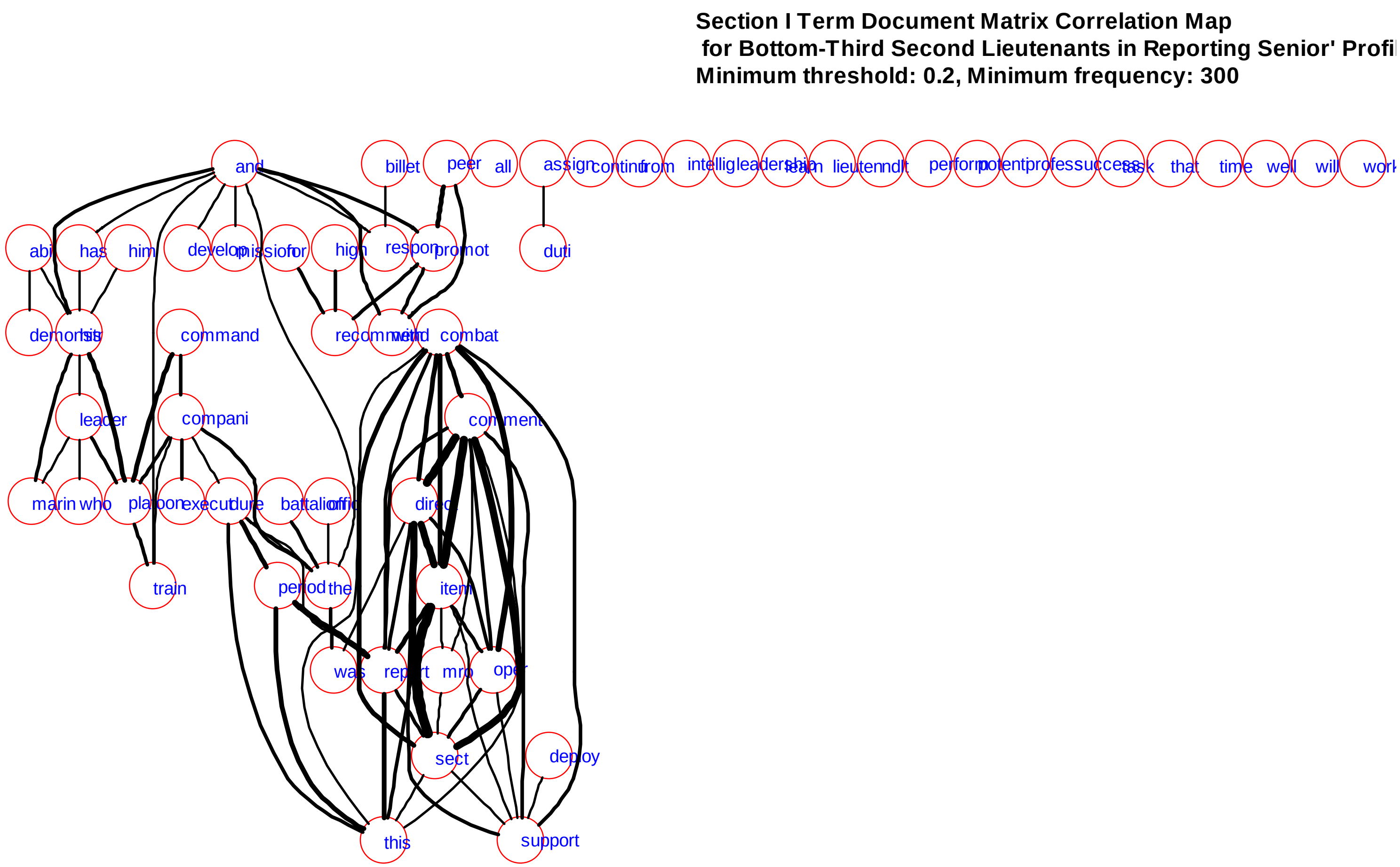


Figure 39. Unsupervised Correlation Map for Bottom Third 2ndtLt Section I

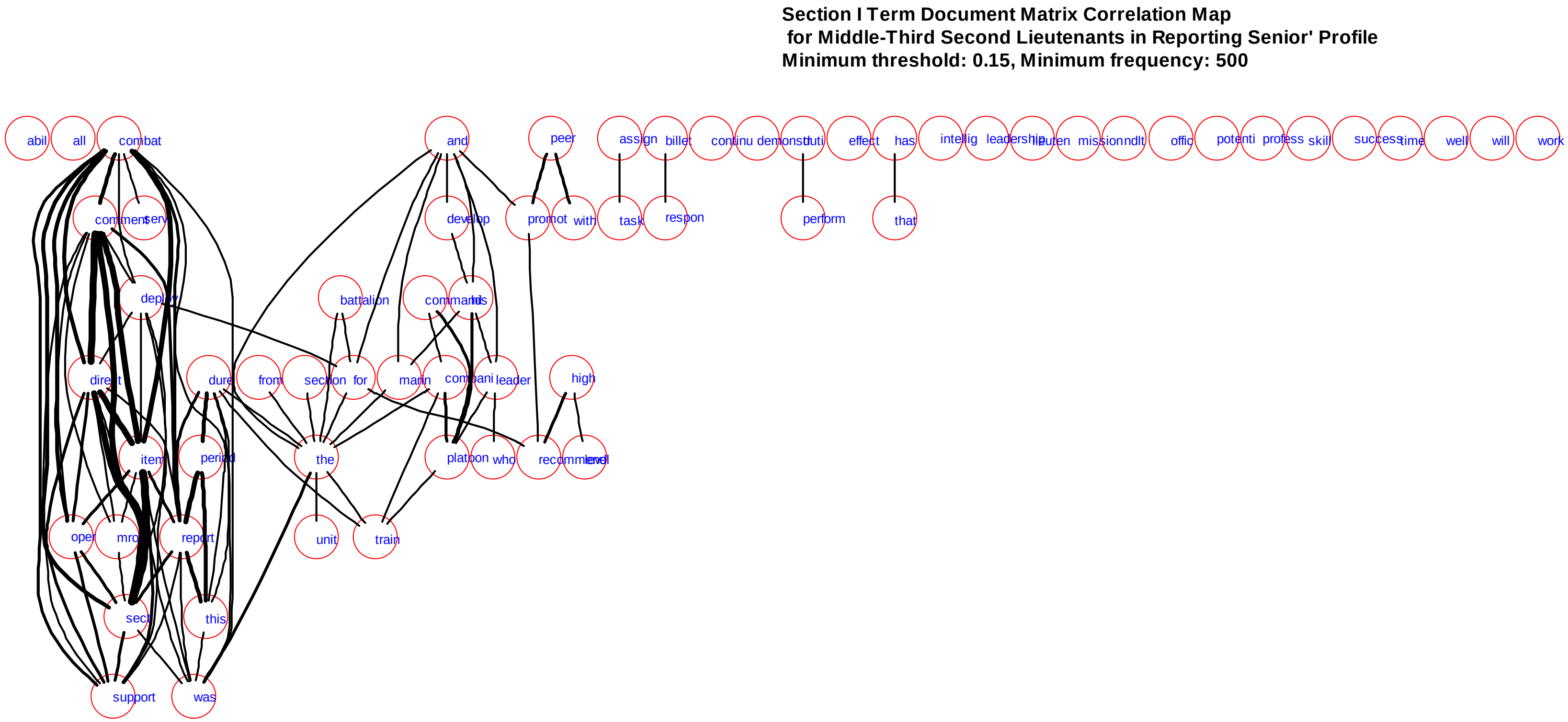


Figure 40. Unsupervised Correlation Map for Middle Third 2ndtLt Section I

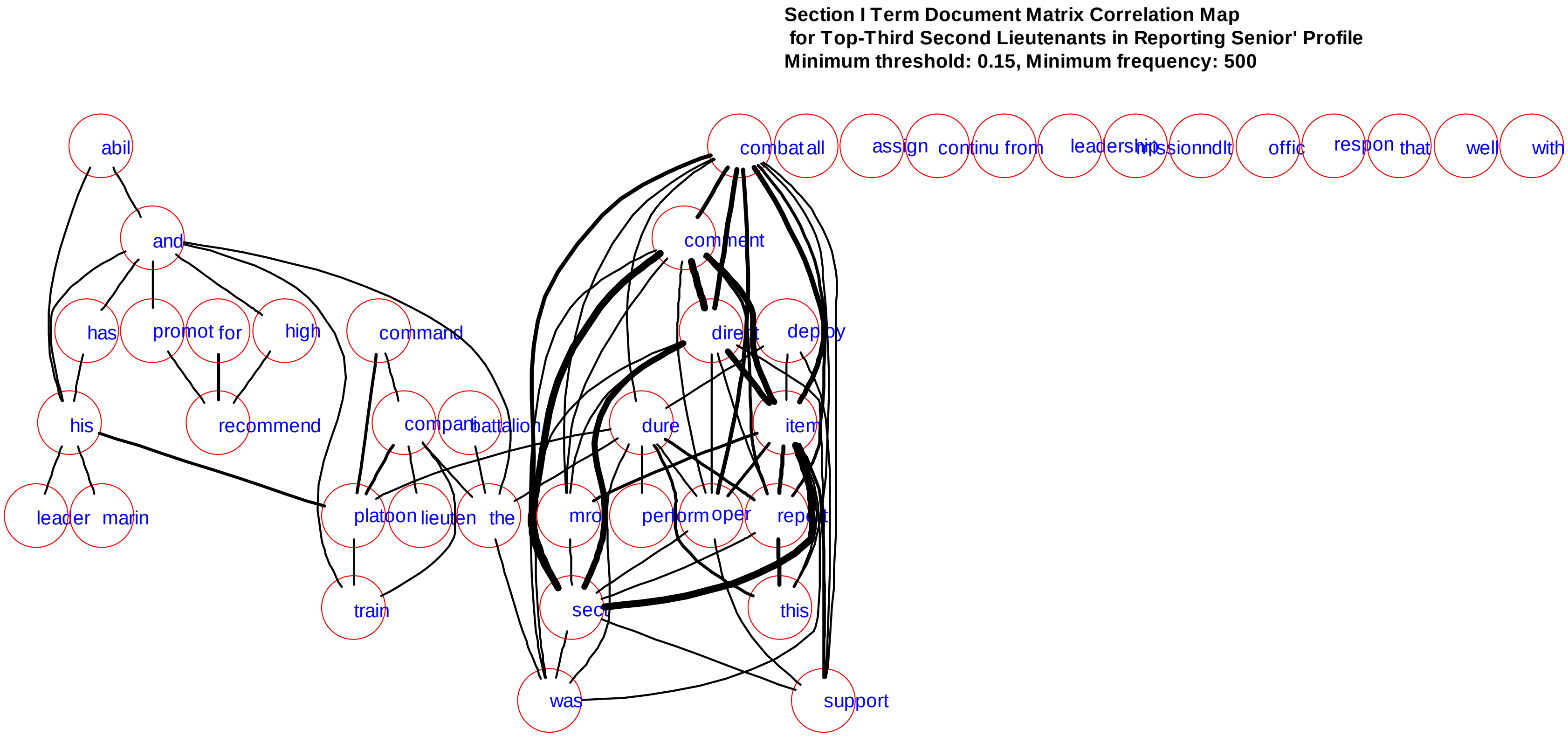


Figure 41. Unsupervised Correlation Map for Top Third 2ndtLt Section I

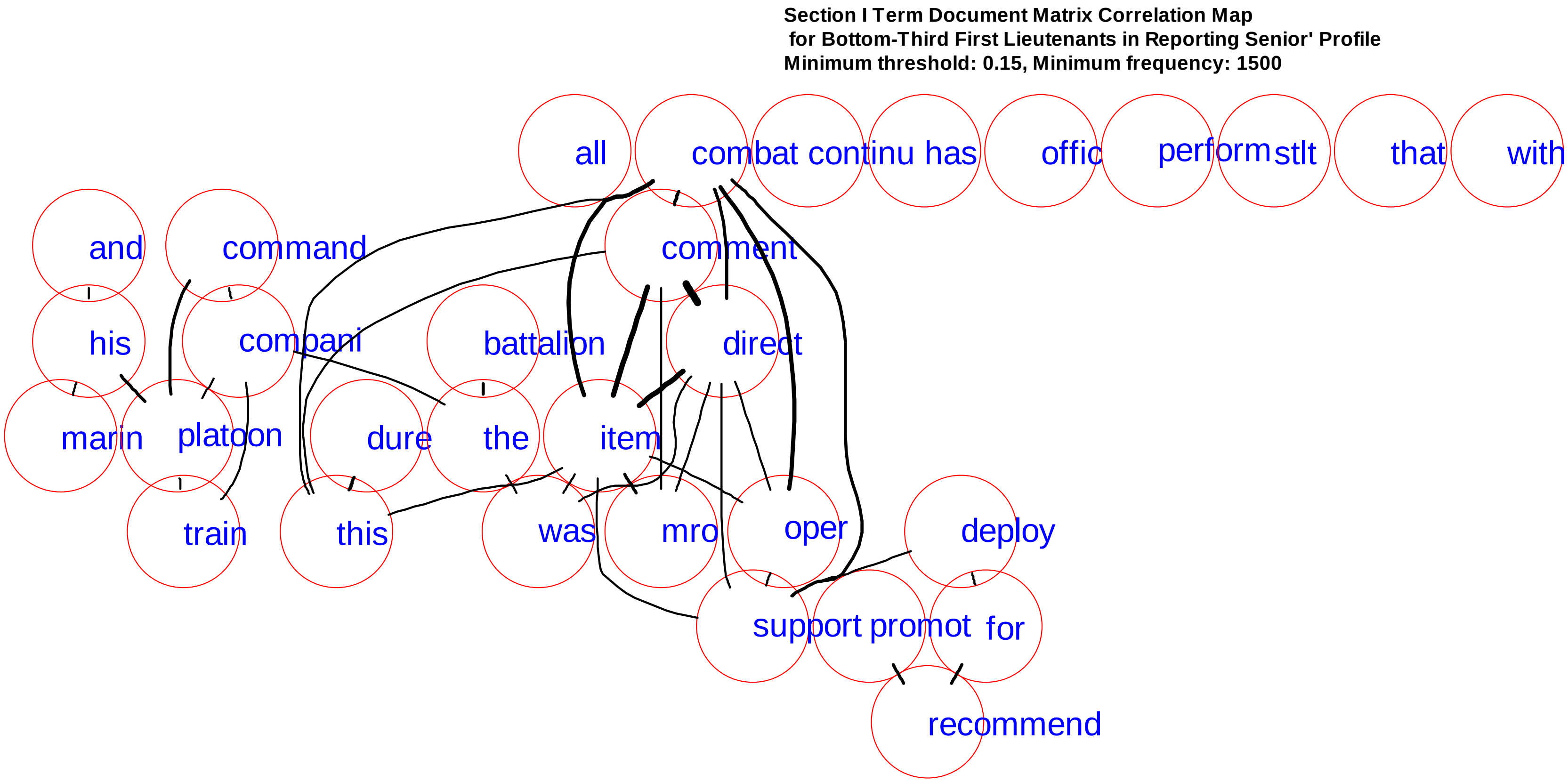


Figure 42. Unsupervised Correlation Map for Bottom Third 1stLt Section I

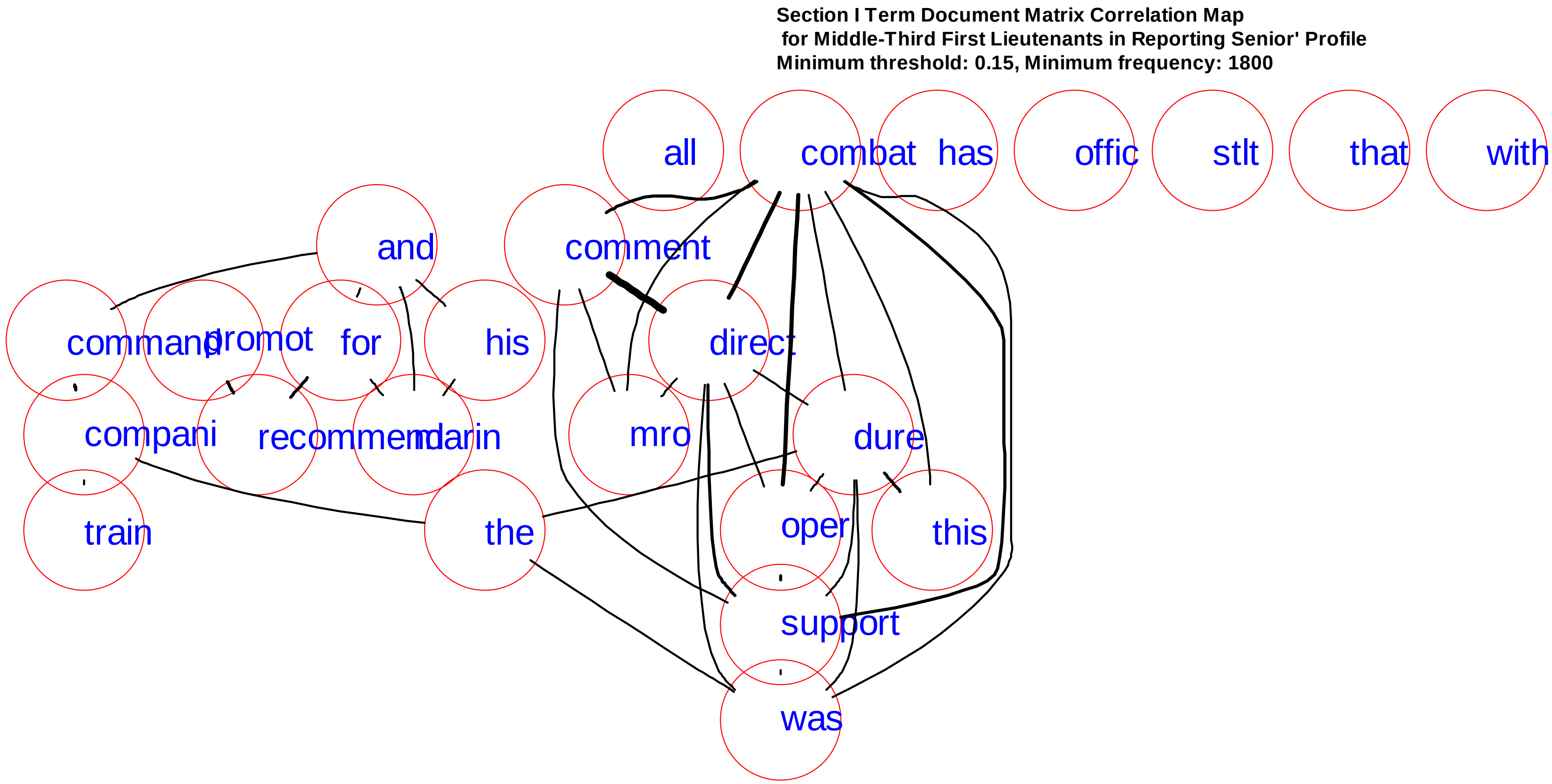


Figure 43. Unsupervised Correlation Map for Middle Third 1stLt Section I

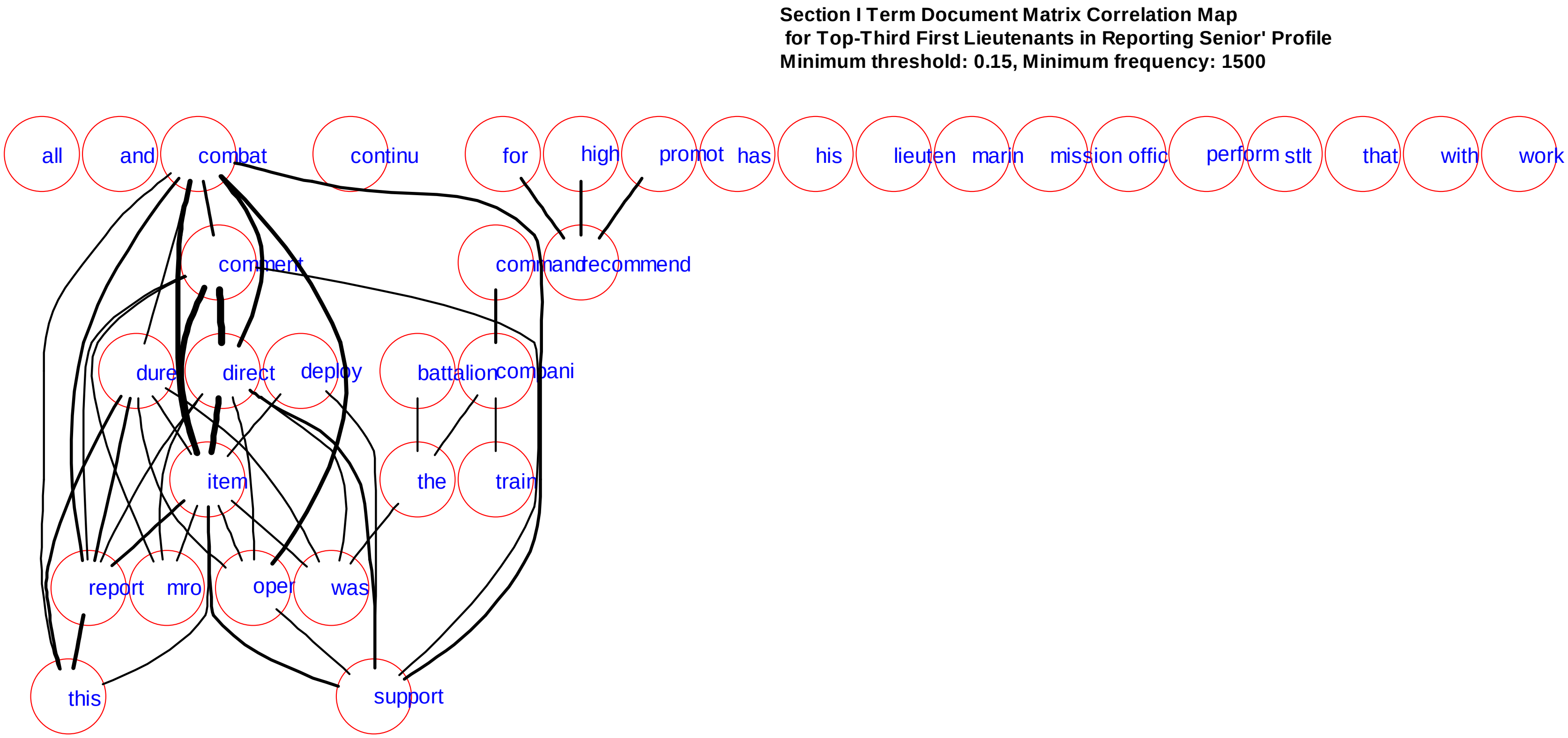


Figure 44. Unsupervised Correlation Map for Top Third 1stLt Section I

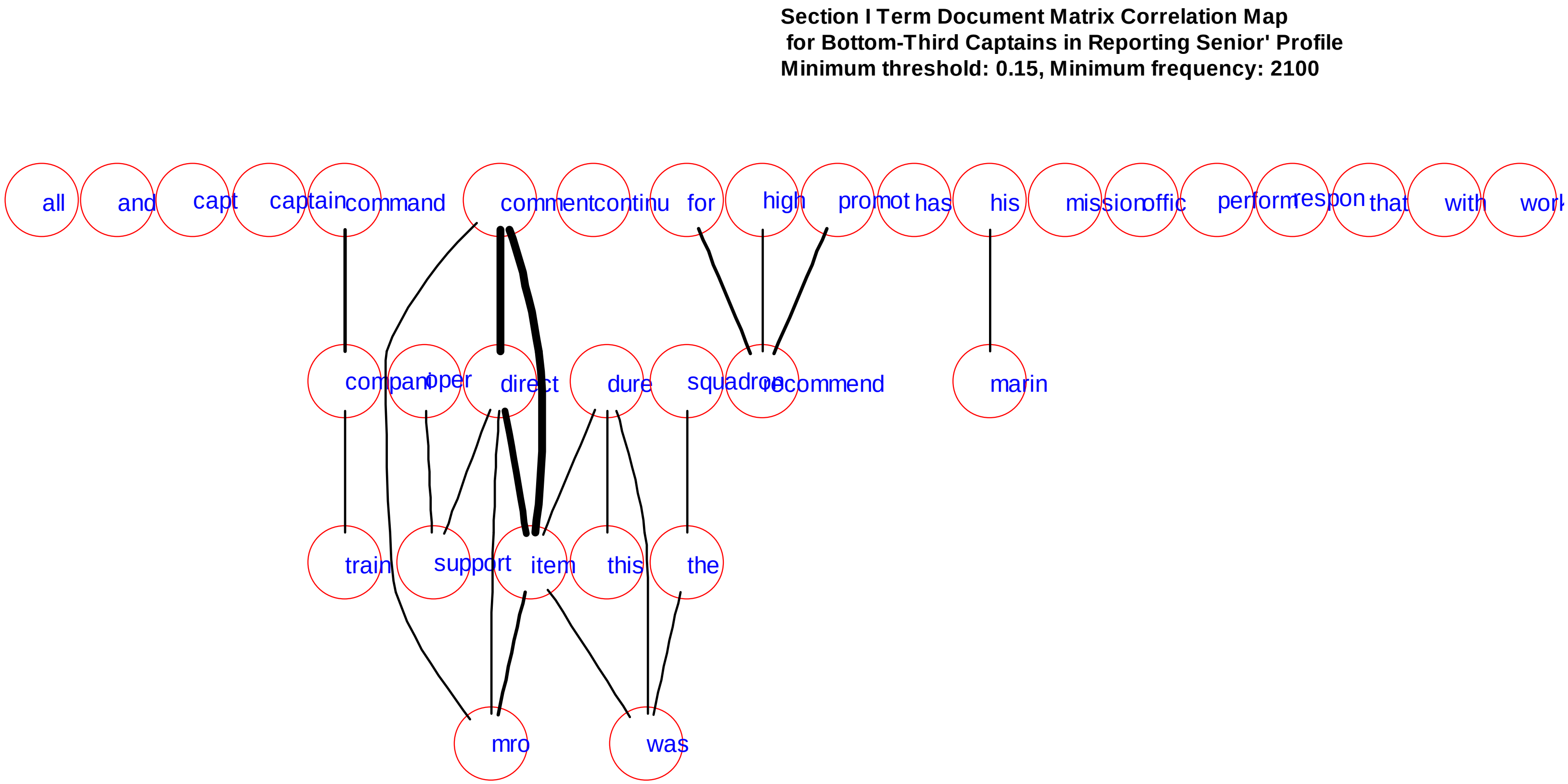


Figure 45. Unsupervised Correlation Map for Bottom Third Capt Section I

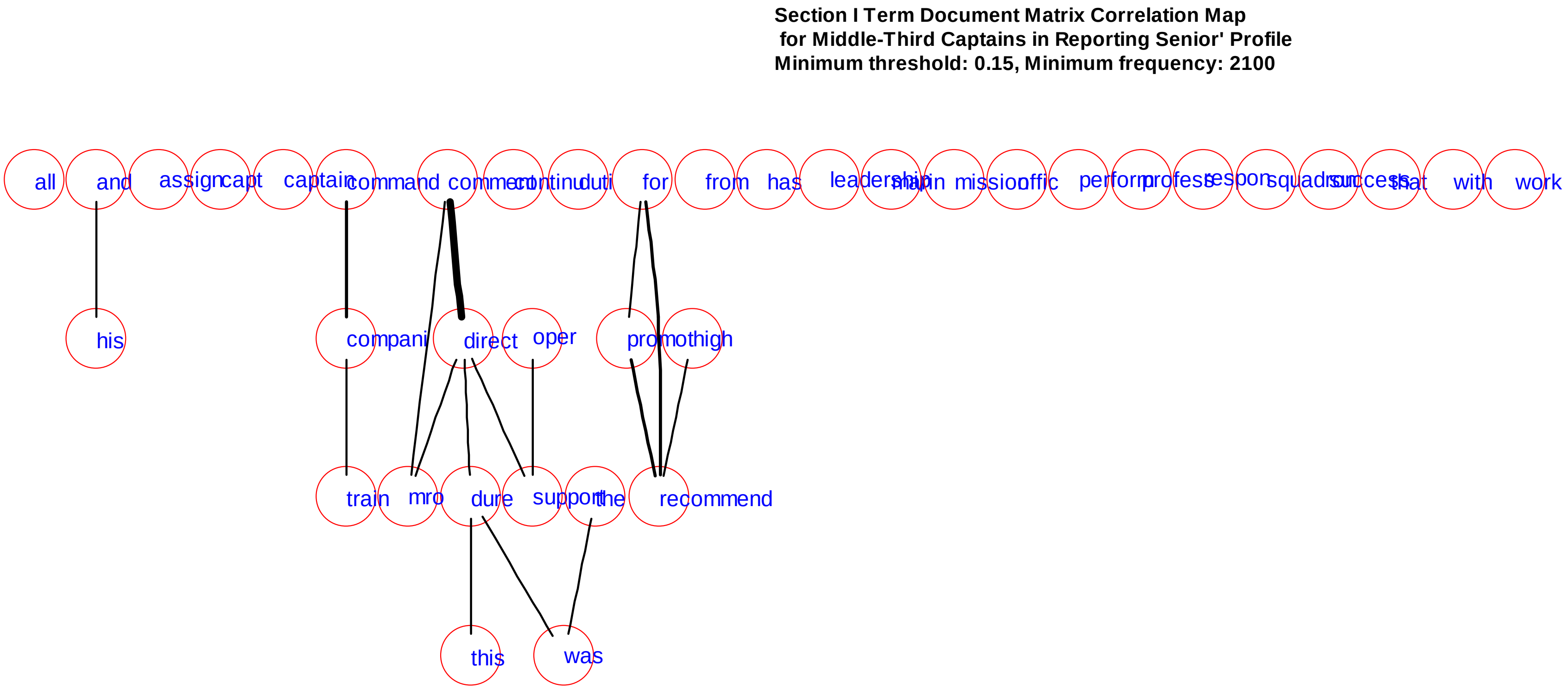


Figure 46. Unsupervised Correlation Map for Middle Third Capt Section I

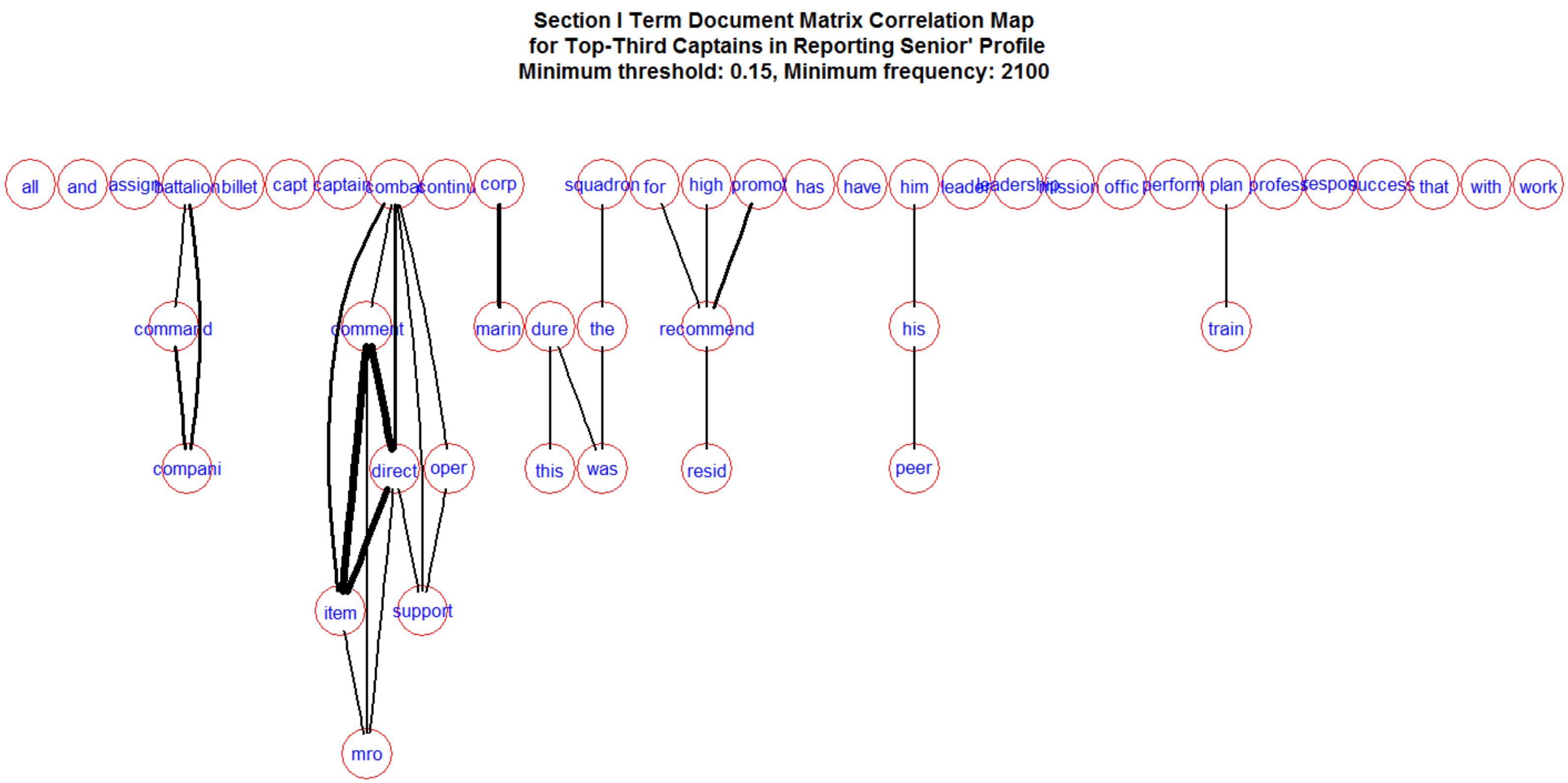


Figure 47. Unsupervised Correlation Map for Top Third Capt Section I

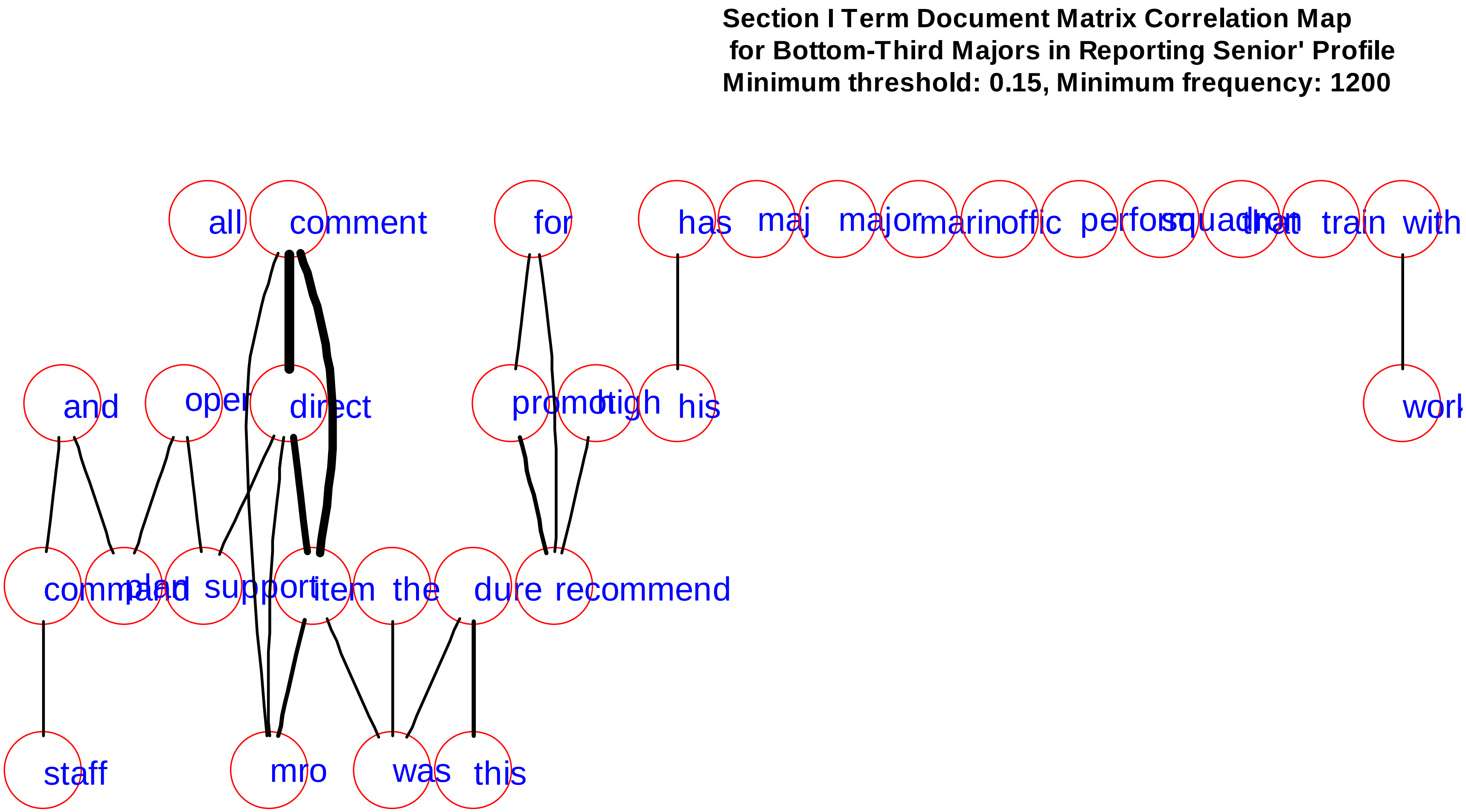


Figure 48. Unsupervised Correlation Map for Bottom Third Maj Section I

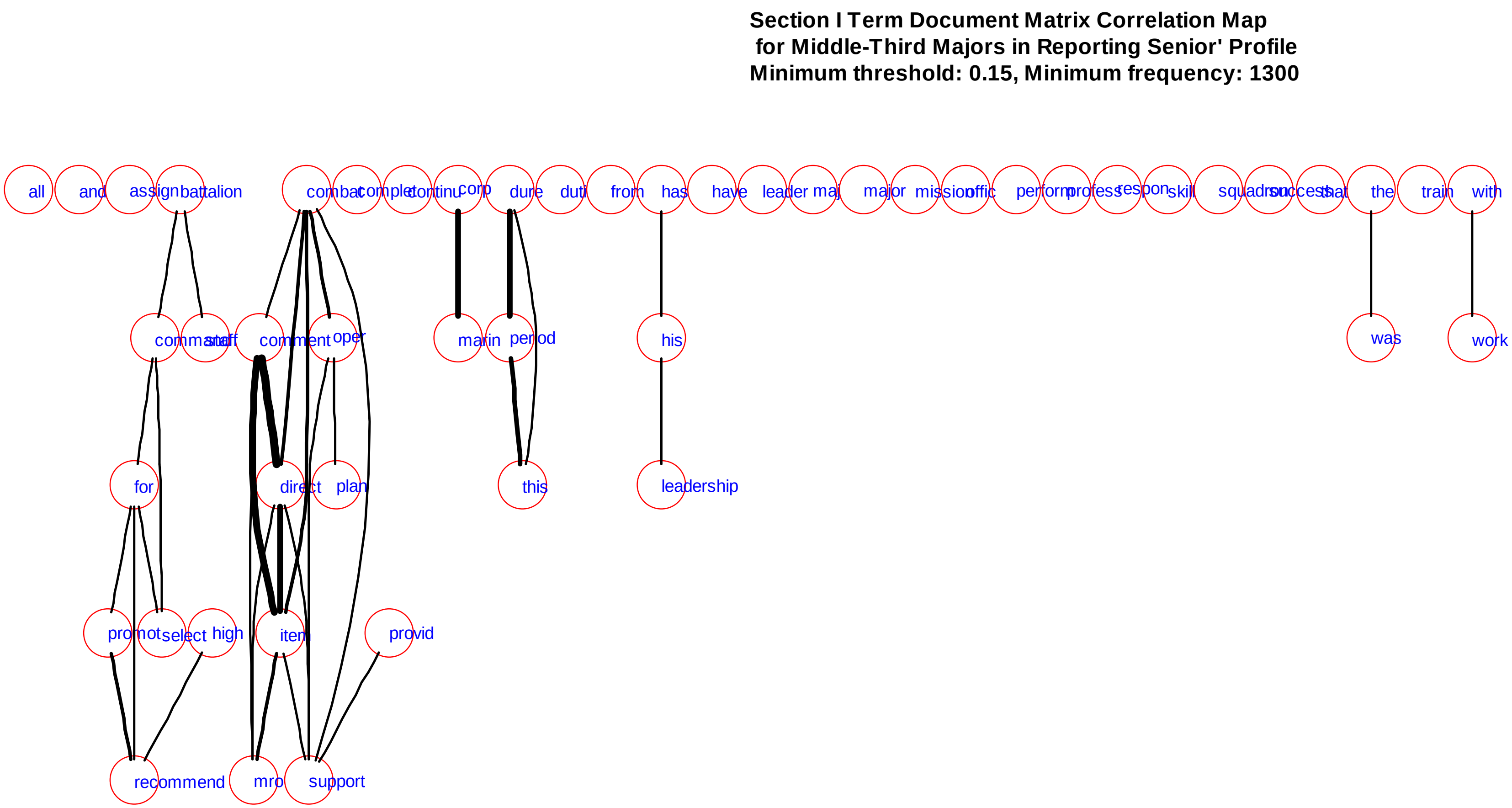


Figure 49. Unsupervised Correlation Map for Middle Third Maj Section I

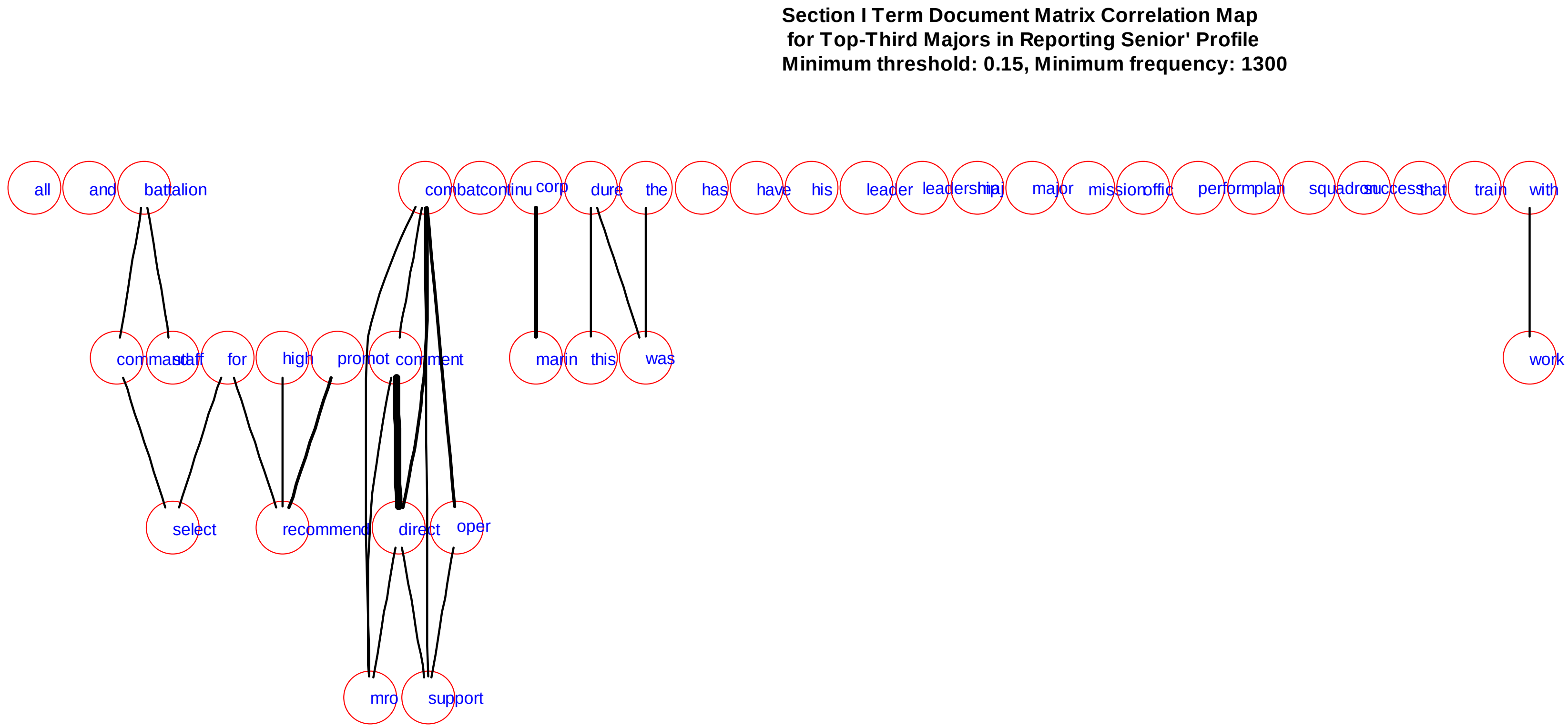


Figure 50. Unsupervised Correlation Map for Top Third Maj Section I

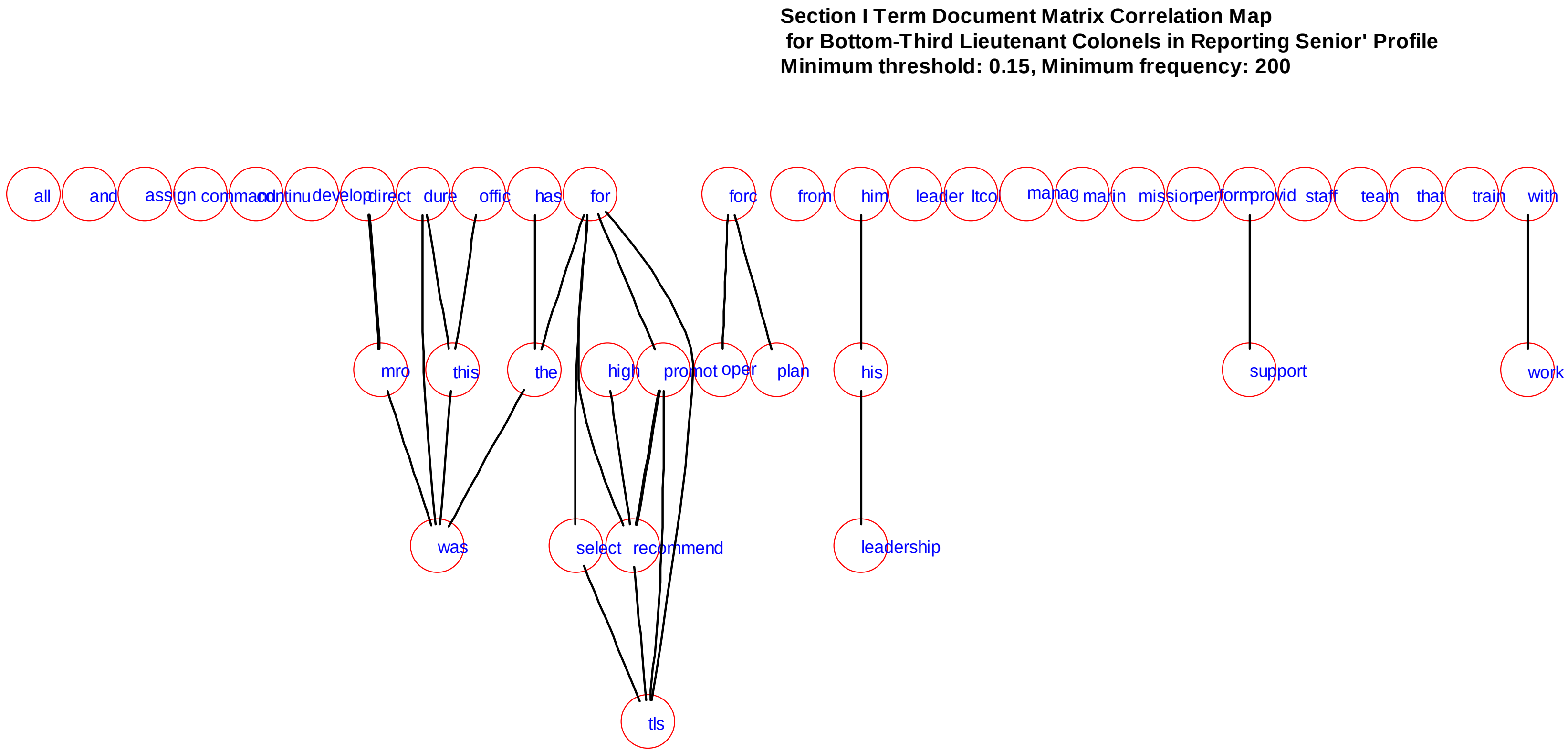


Figure 51. Unsupervised Correlation Map for Bottom Third LtCol Section I

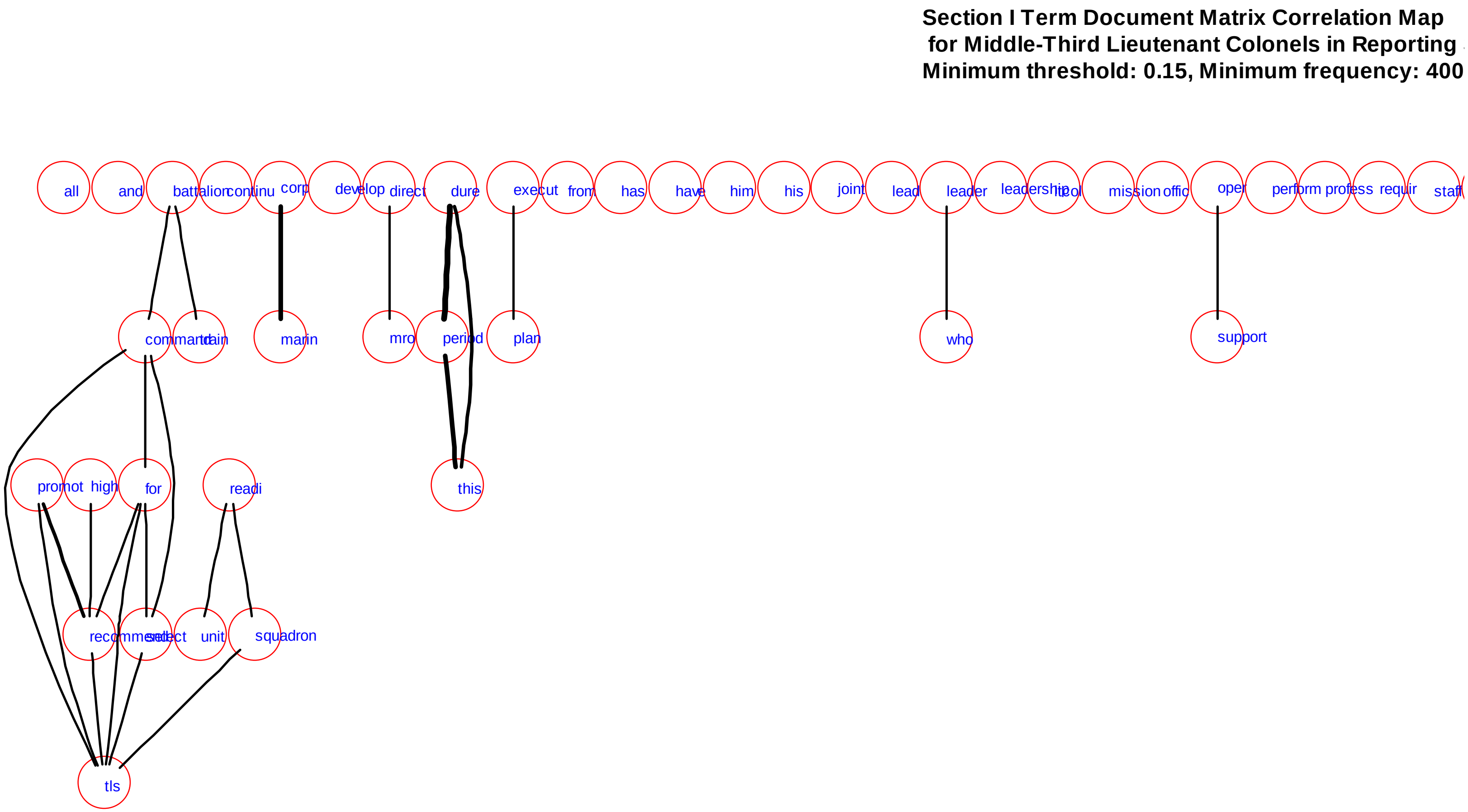


Figure 52. Unsupervised Correlation Map for Middle Third LtCol Section I

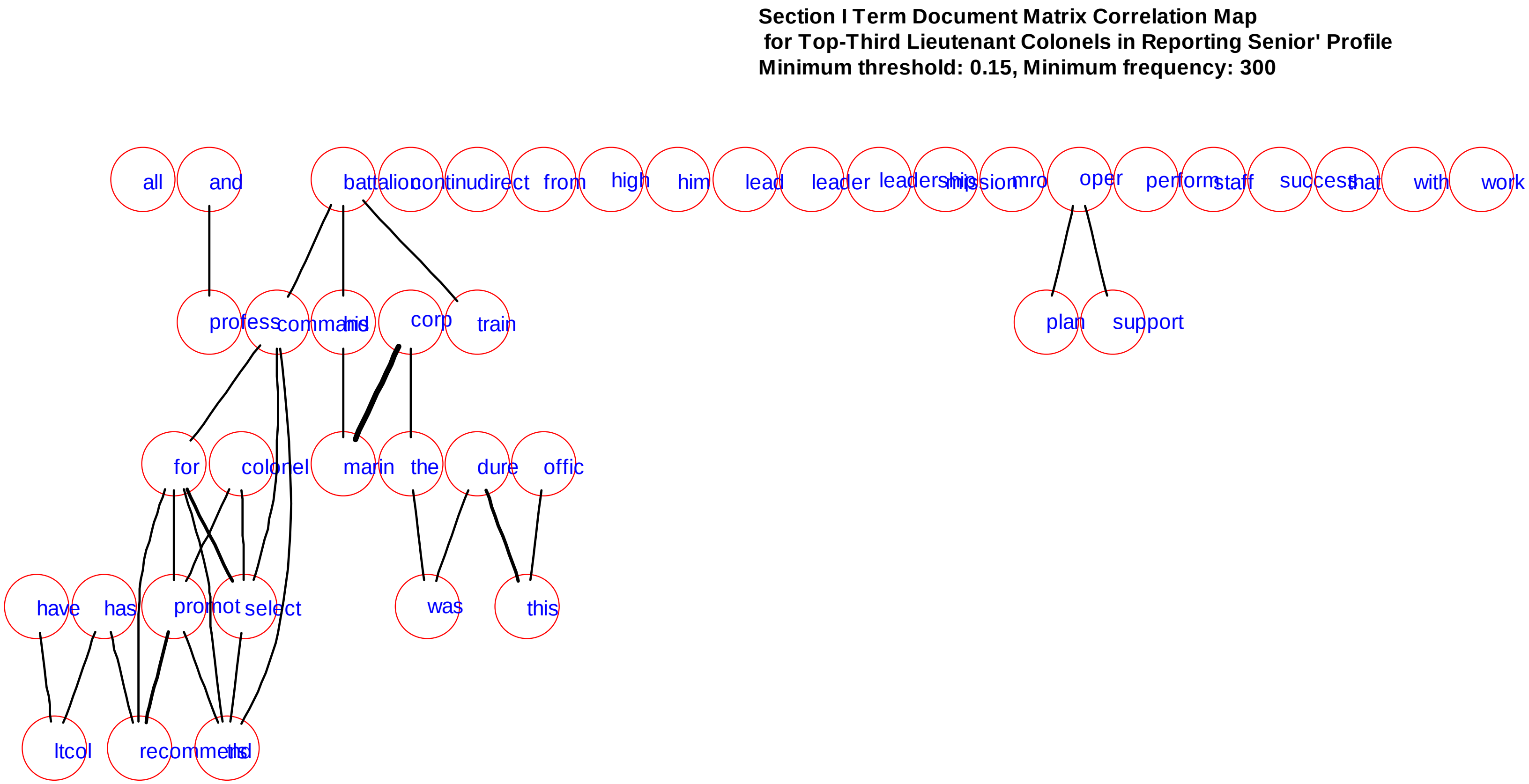


Figure 53. Unsupervised Correlation Map for Top Third LtCol Section I

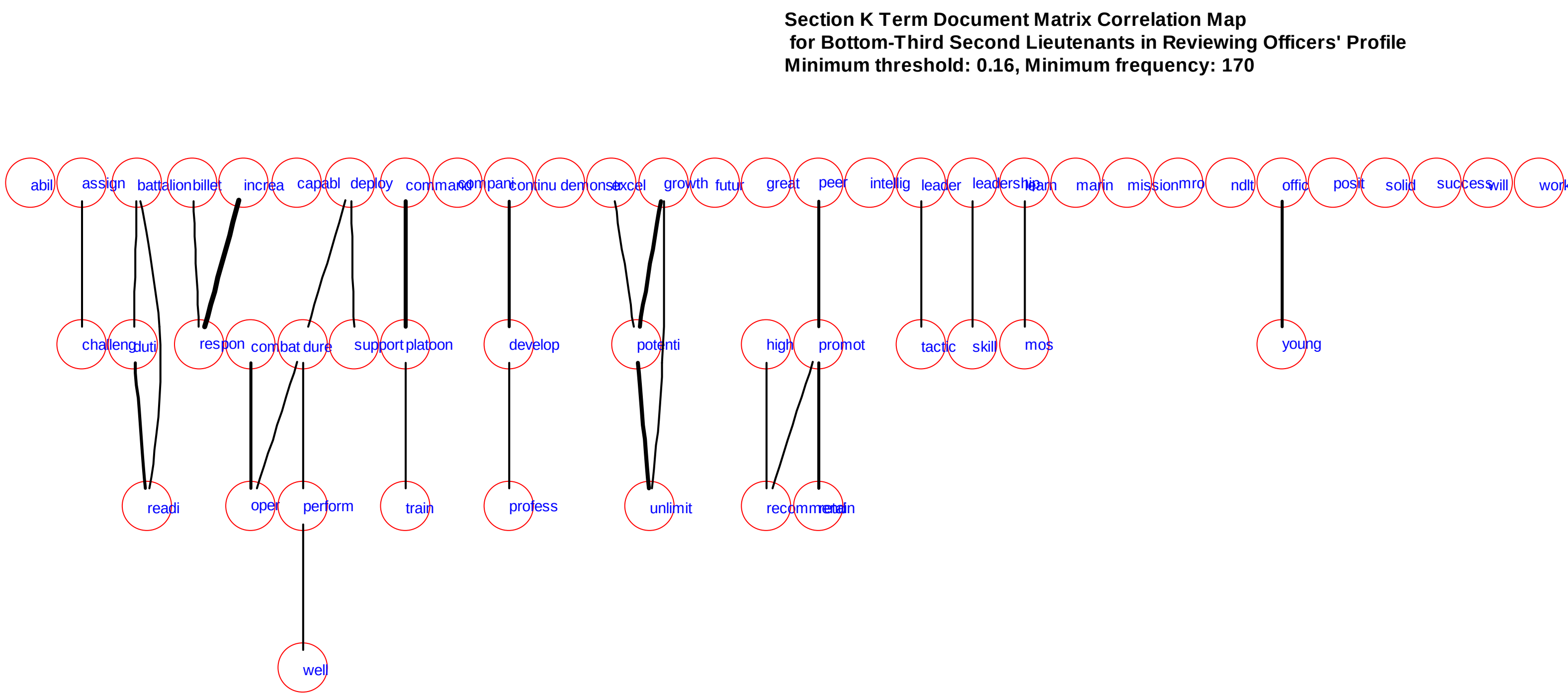
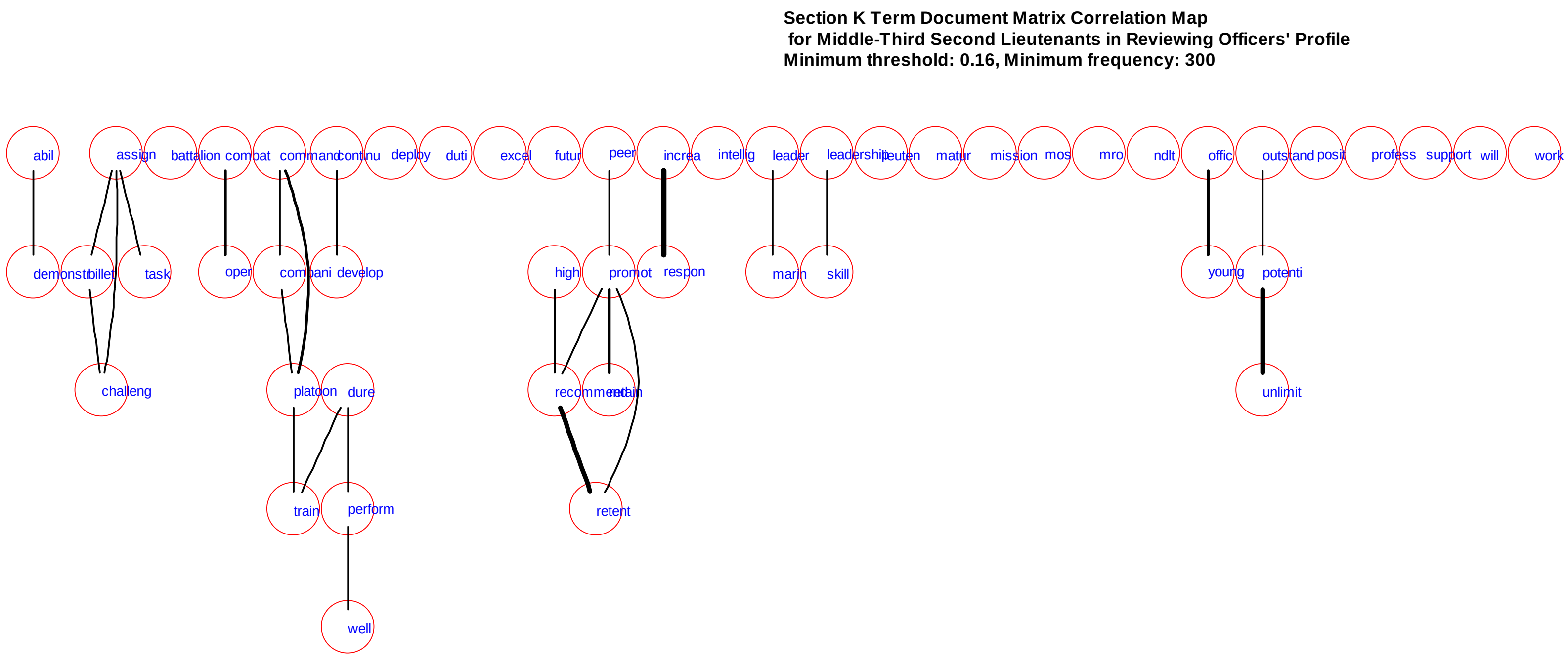


Figure 54. Unsupervised Correlation Map for Bottom Third 2ndLt Section K



`

Figure 55. Unsupervised Correlation Map for Middle Third 2ndLt Section K

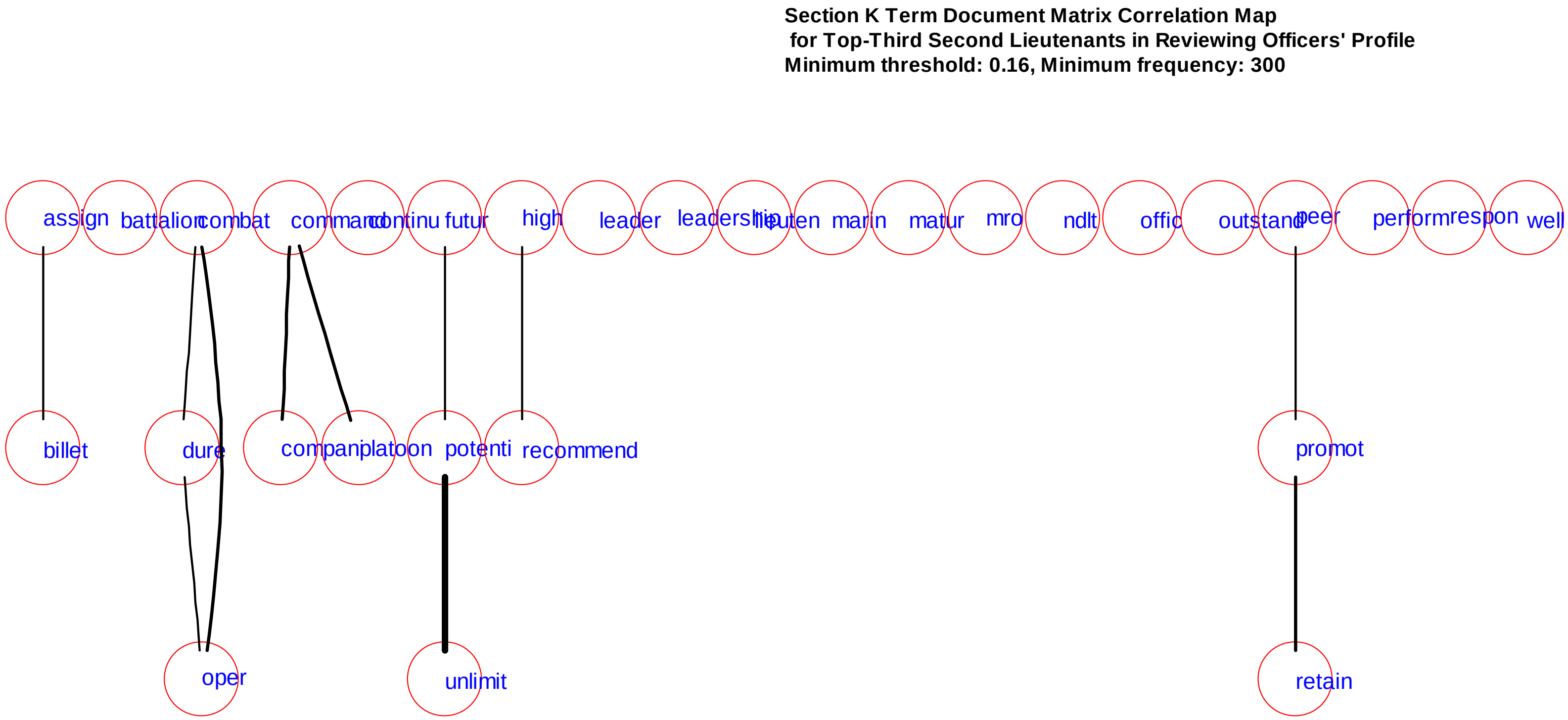


Figure 56. Unsupervised Correlation Map for Top Third 2ndLt Section K

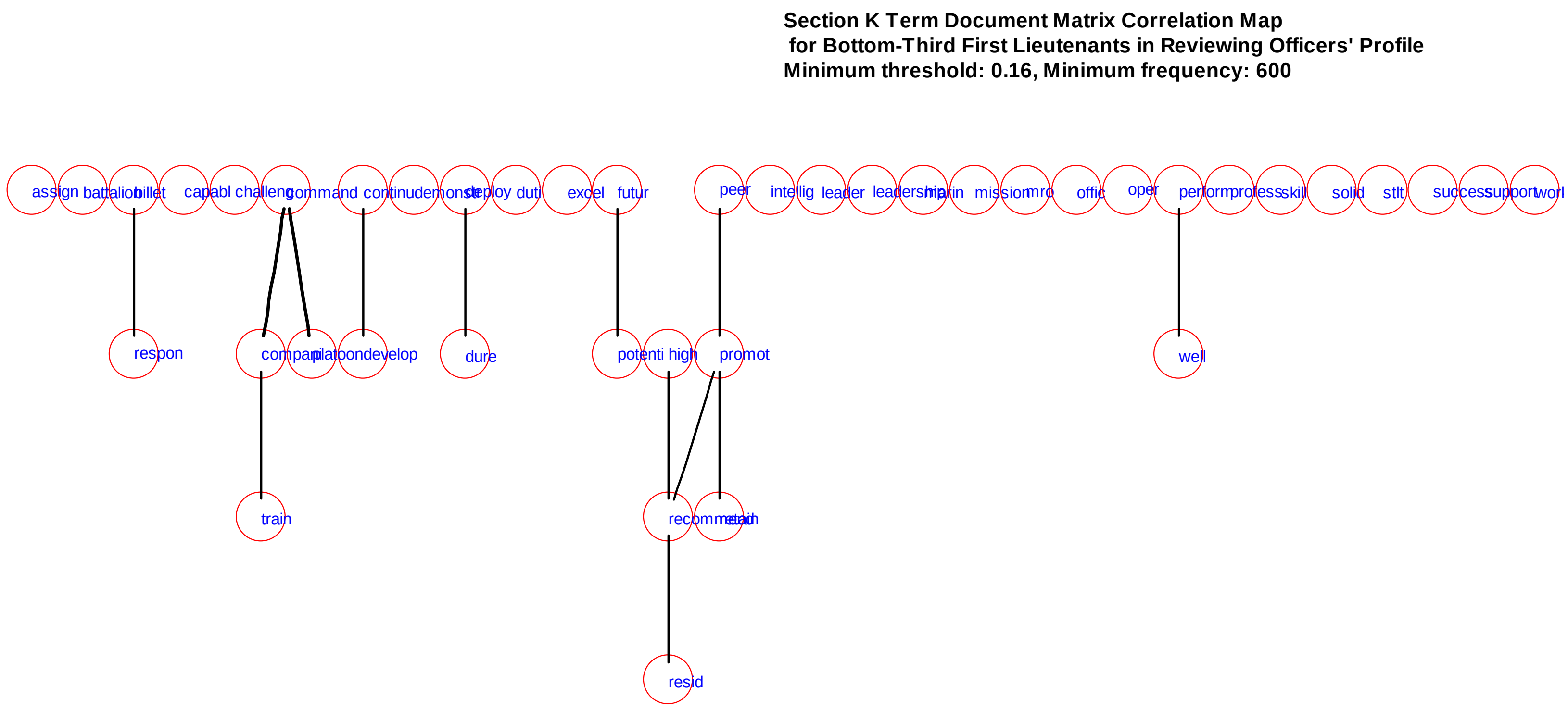


Figure 57. Unsupervised Correlation Map for Bottom Third 1stLt Section K

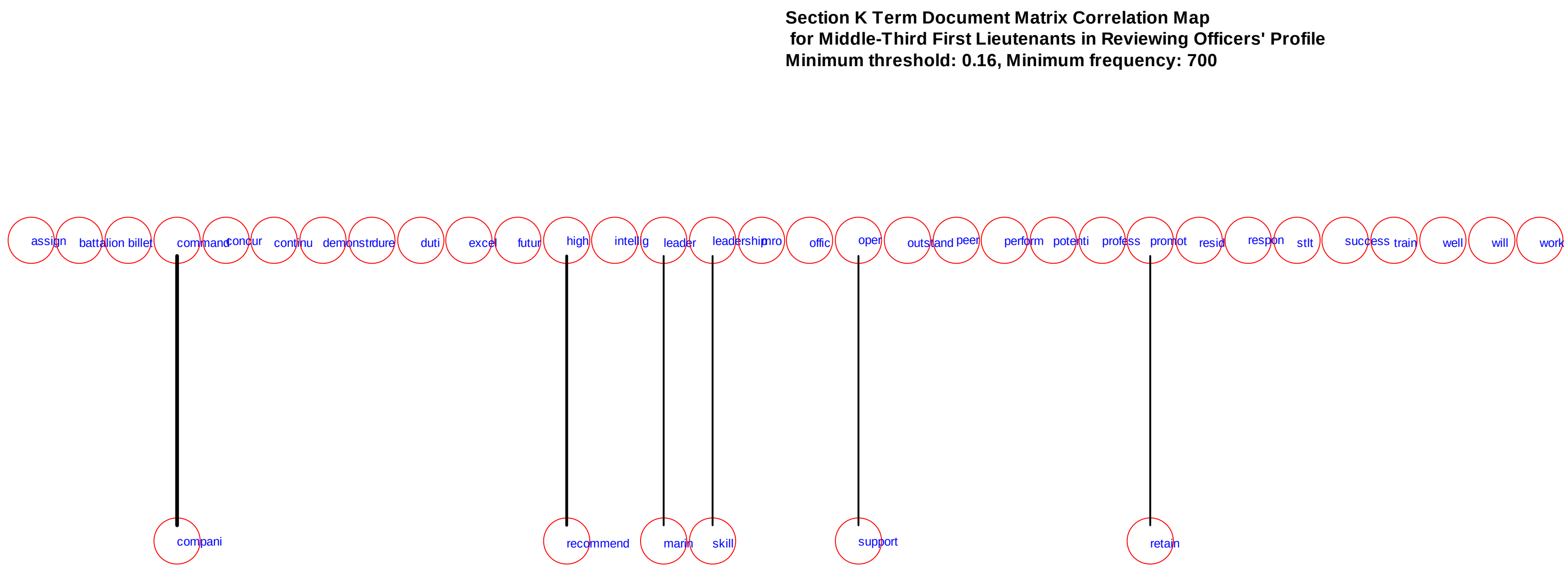


Figure 58. Unsupervised Correlation Map for Middle Third 1stLt Section K

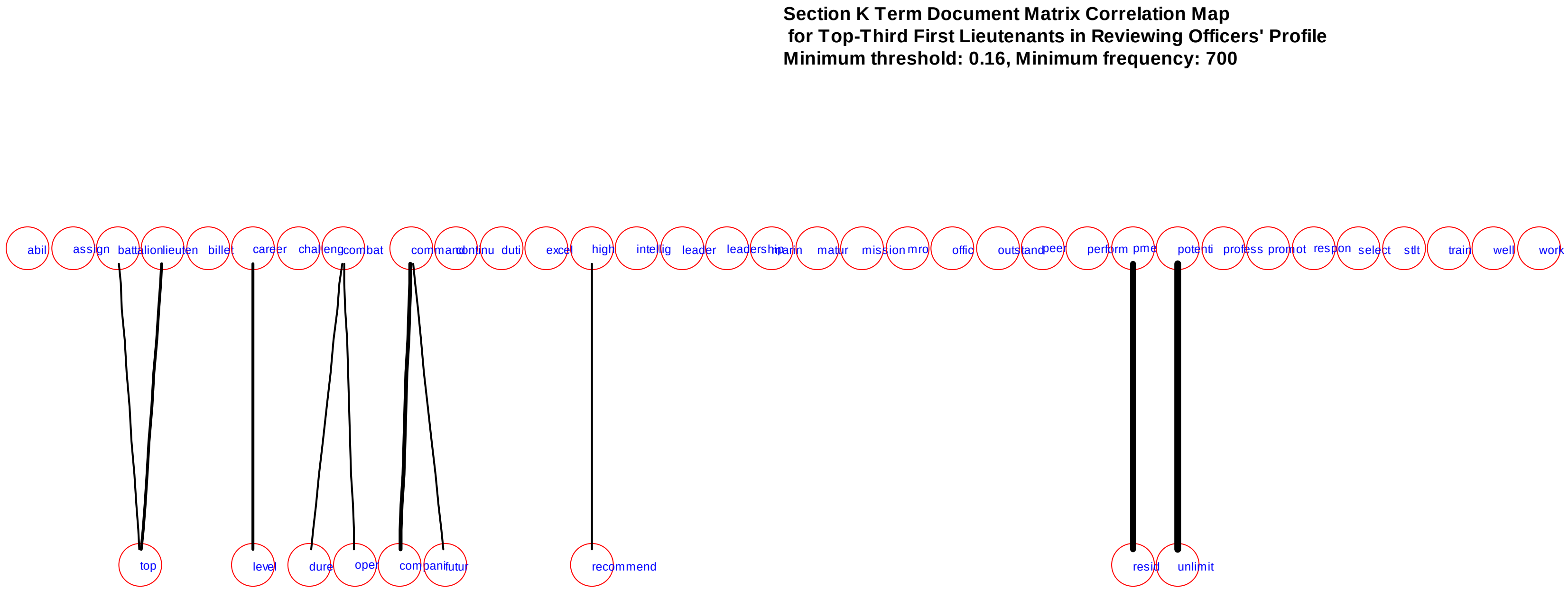


Figure 59. Unsupervised Correlation Map for Top Third 1stLt Section K

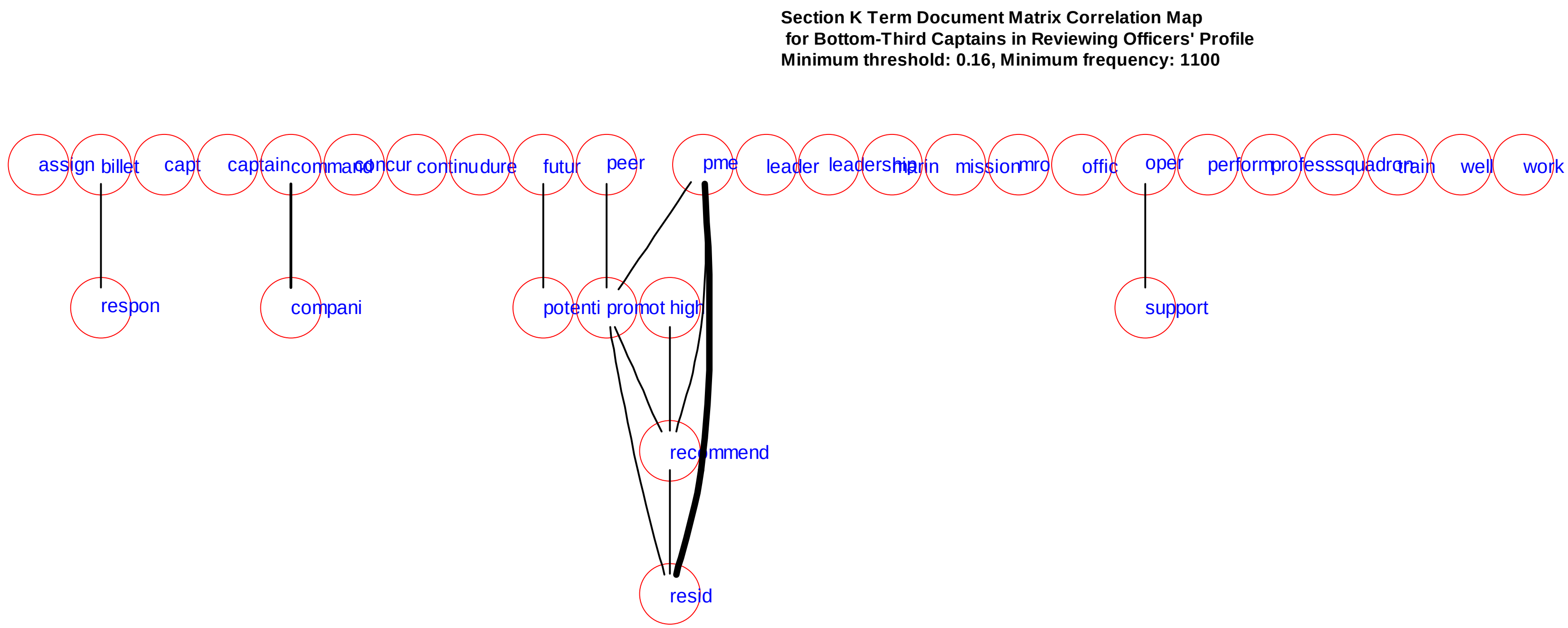


Figure 60. Unsupervised Correlation Map for Bottom Third Capt Section K

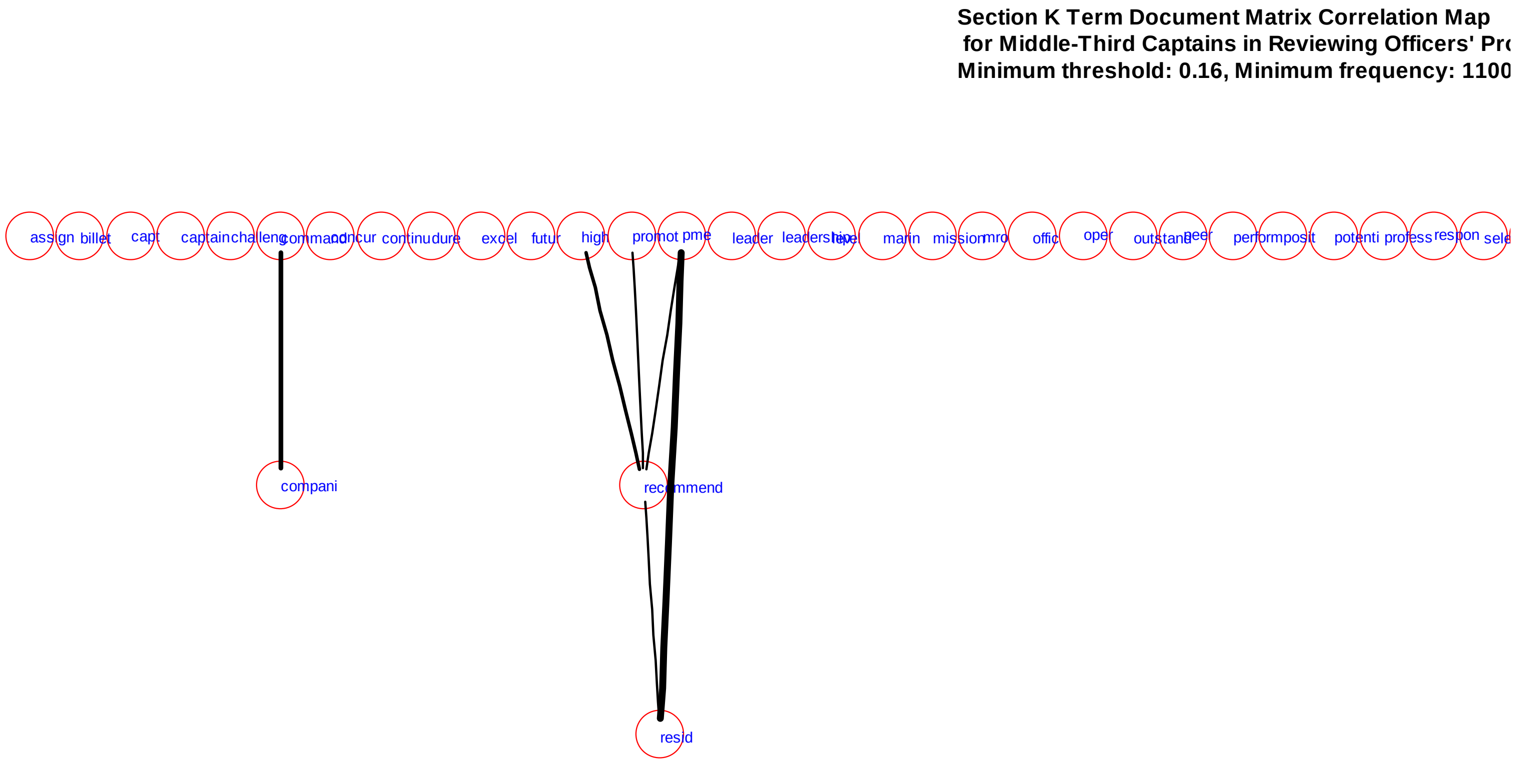


Figure 61. Unsupervised Correlation Map for Middle Third Capt Section K

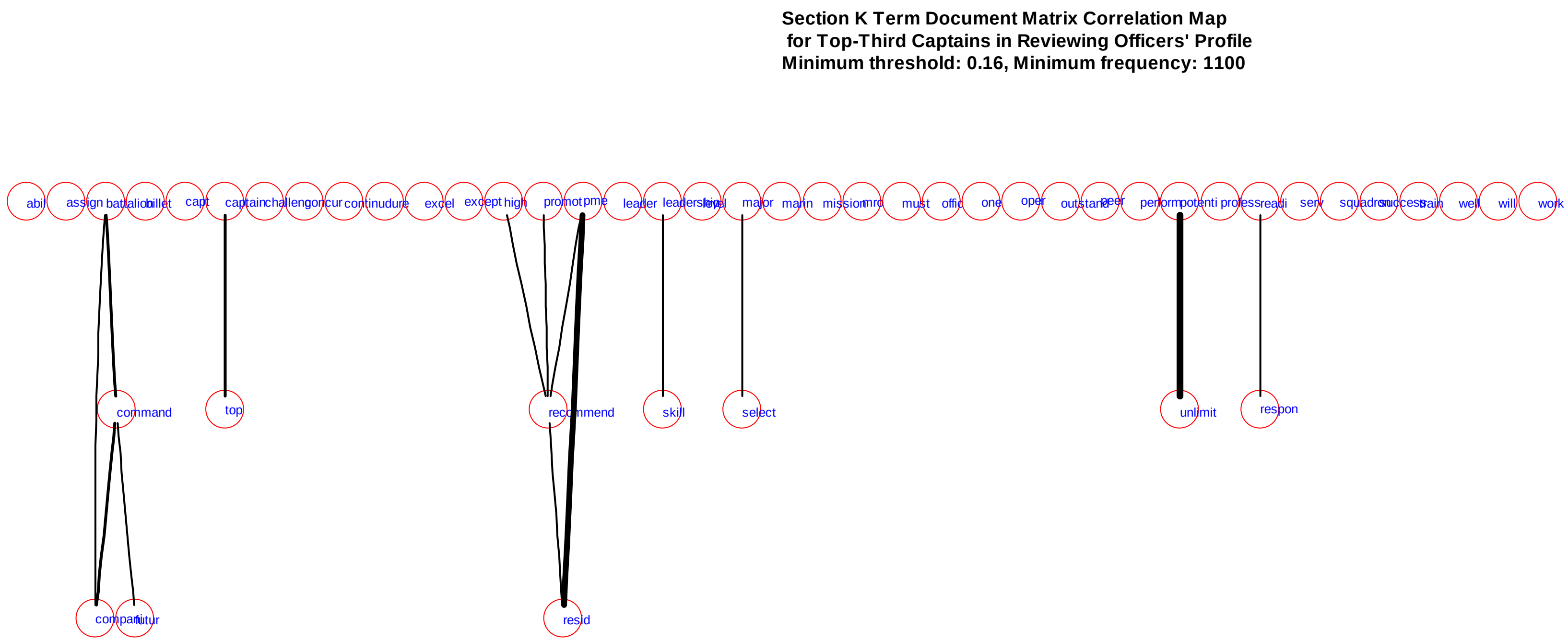


Figure 62. Unsupervised Correlation Map for Top Third Capt Section K

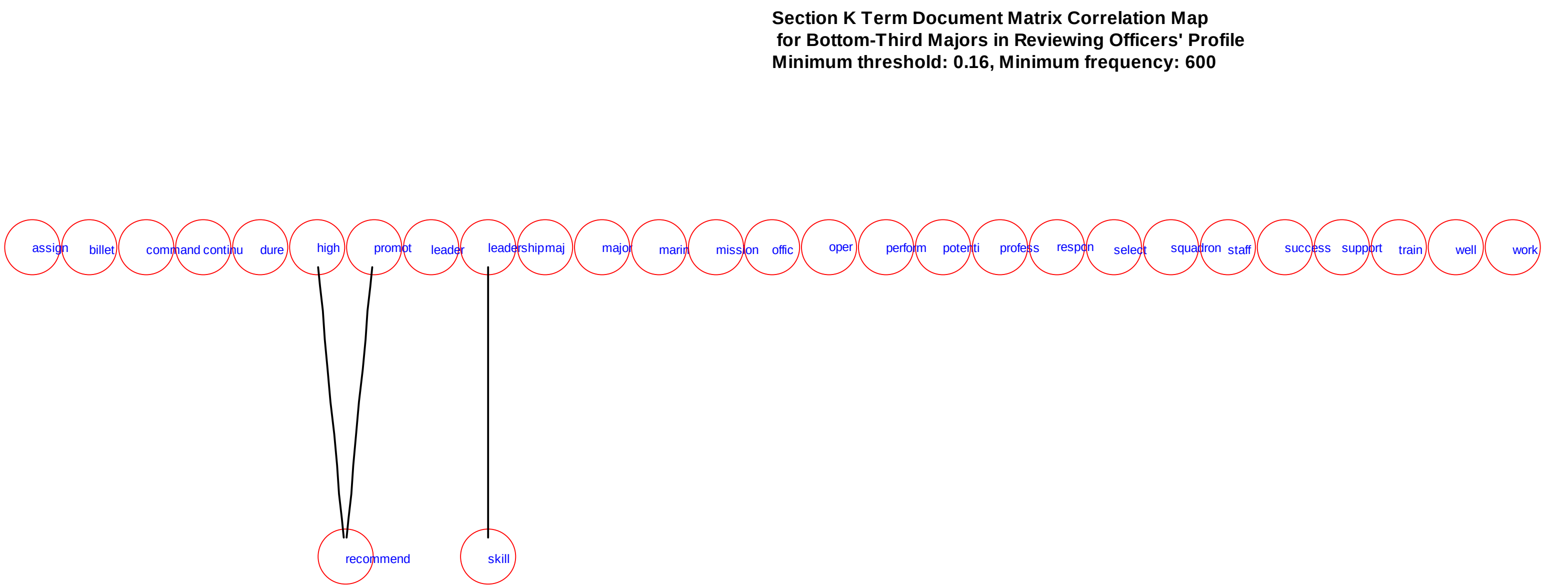


Figure 63. Unsupervised Correlation Map for Bottom Third Maj Section K

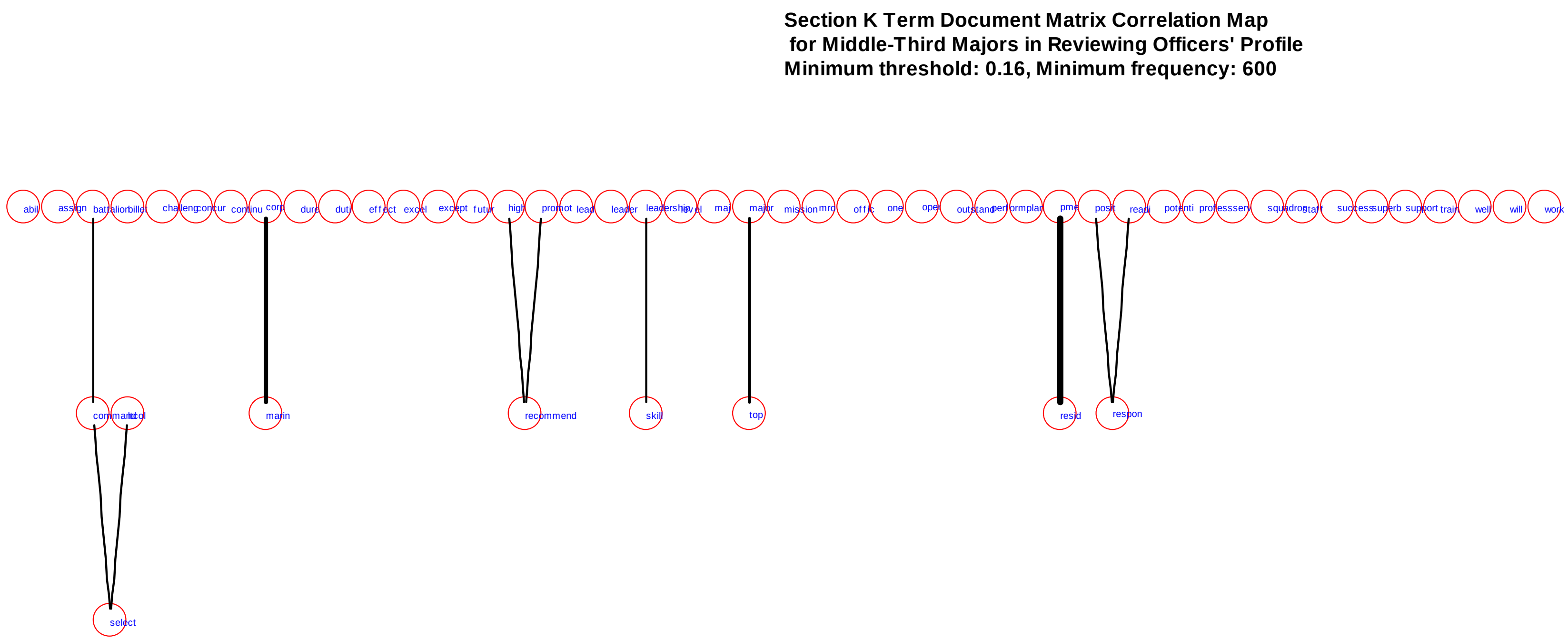


Figure 64. Unsupervised Correlation Map for Middle Third Maj Section K

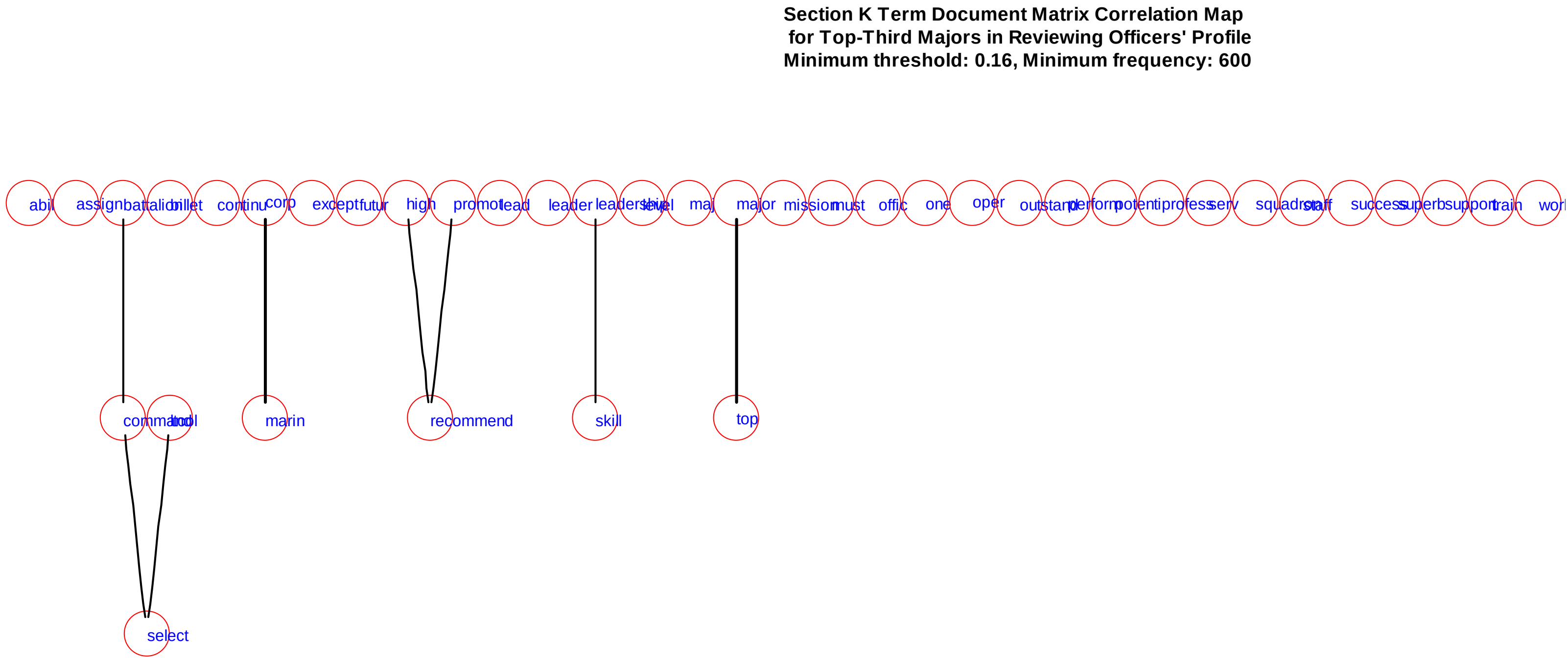


Figure 65. Unsupervised Correlation Map for Top Third Maj Section K

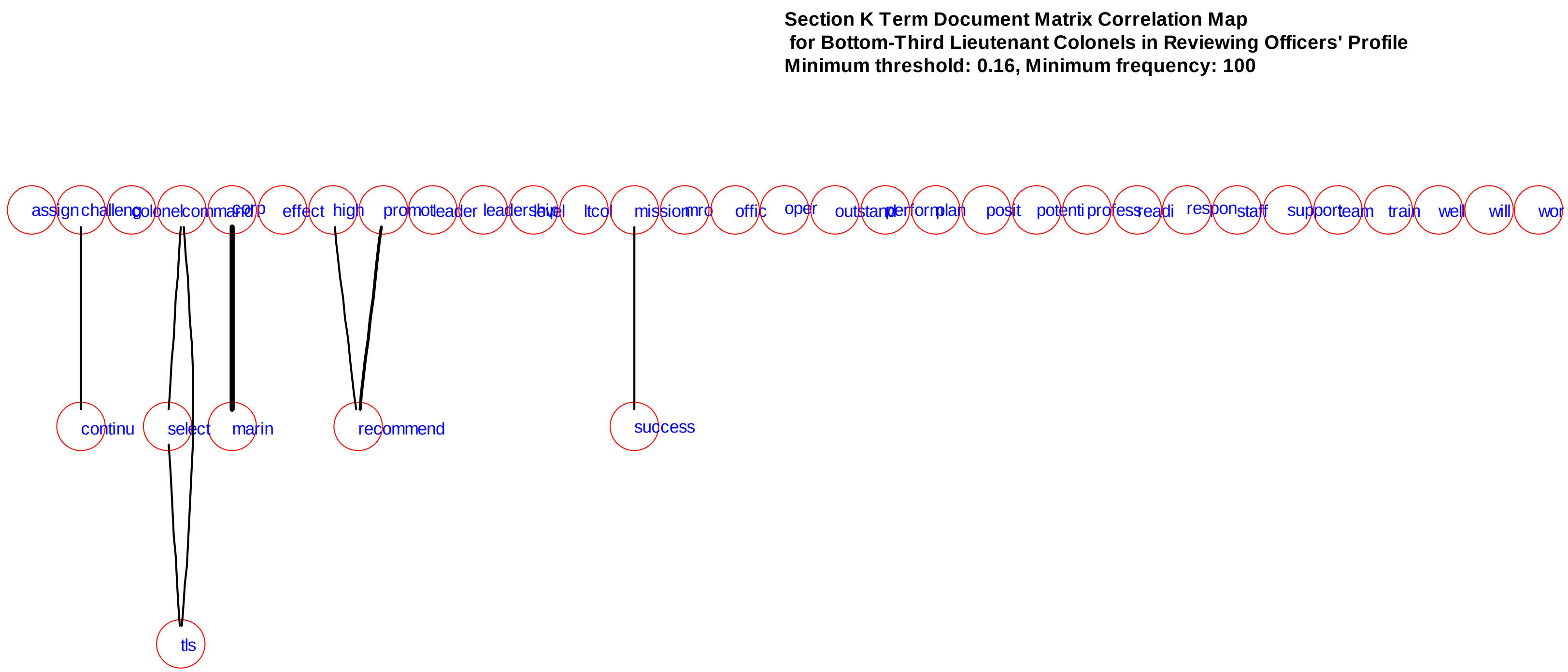


Figure 66. Unsupervised Correlation Map for Bottom Third LtCol Section K

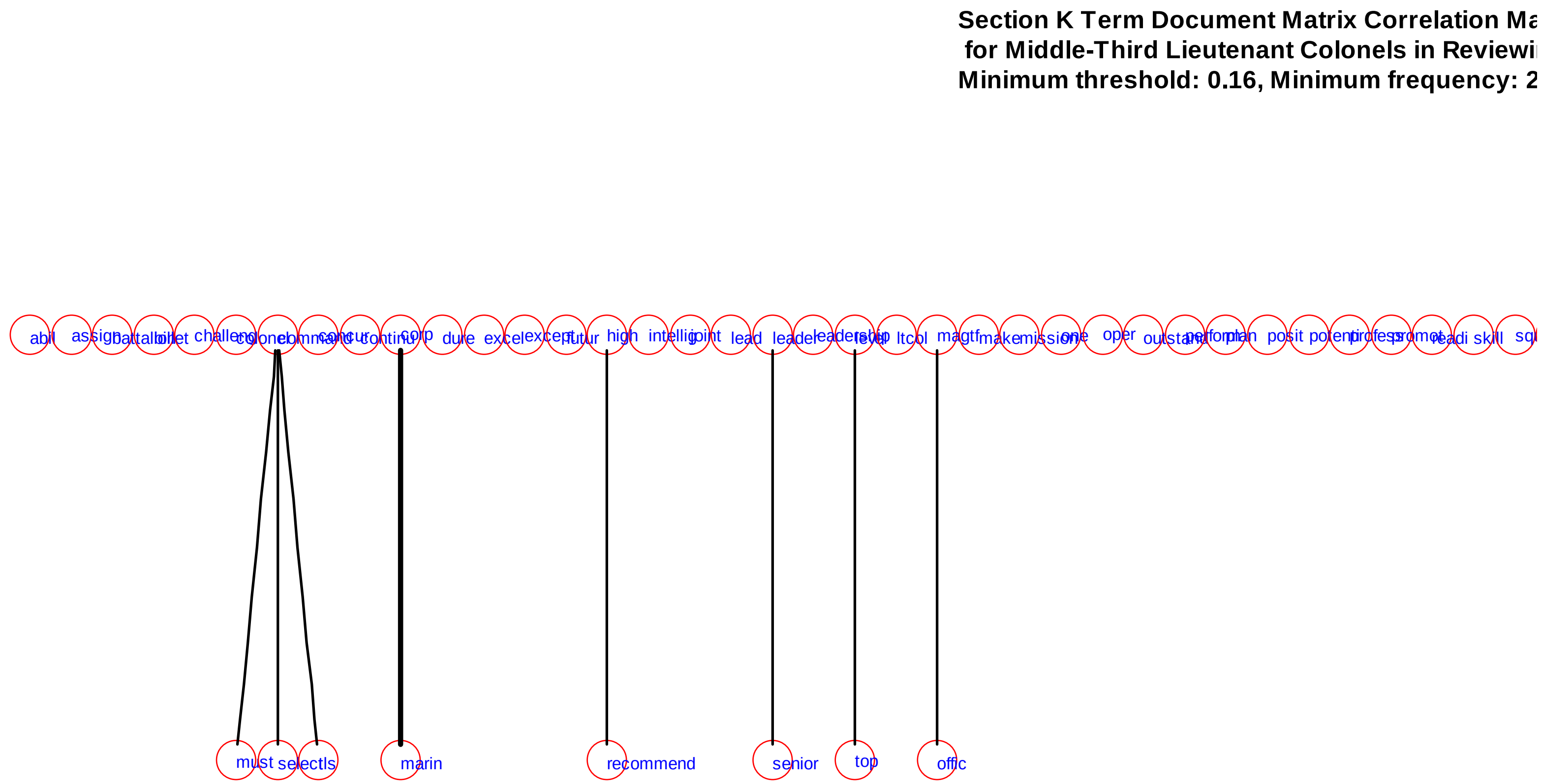


Figure 67. Unsupervised Correlation Map for Middle Third LtCol Section K

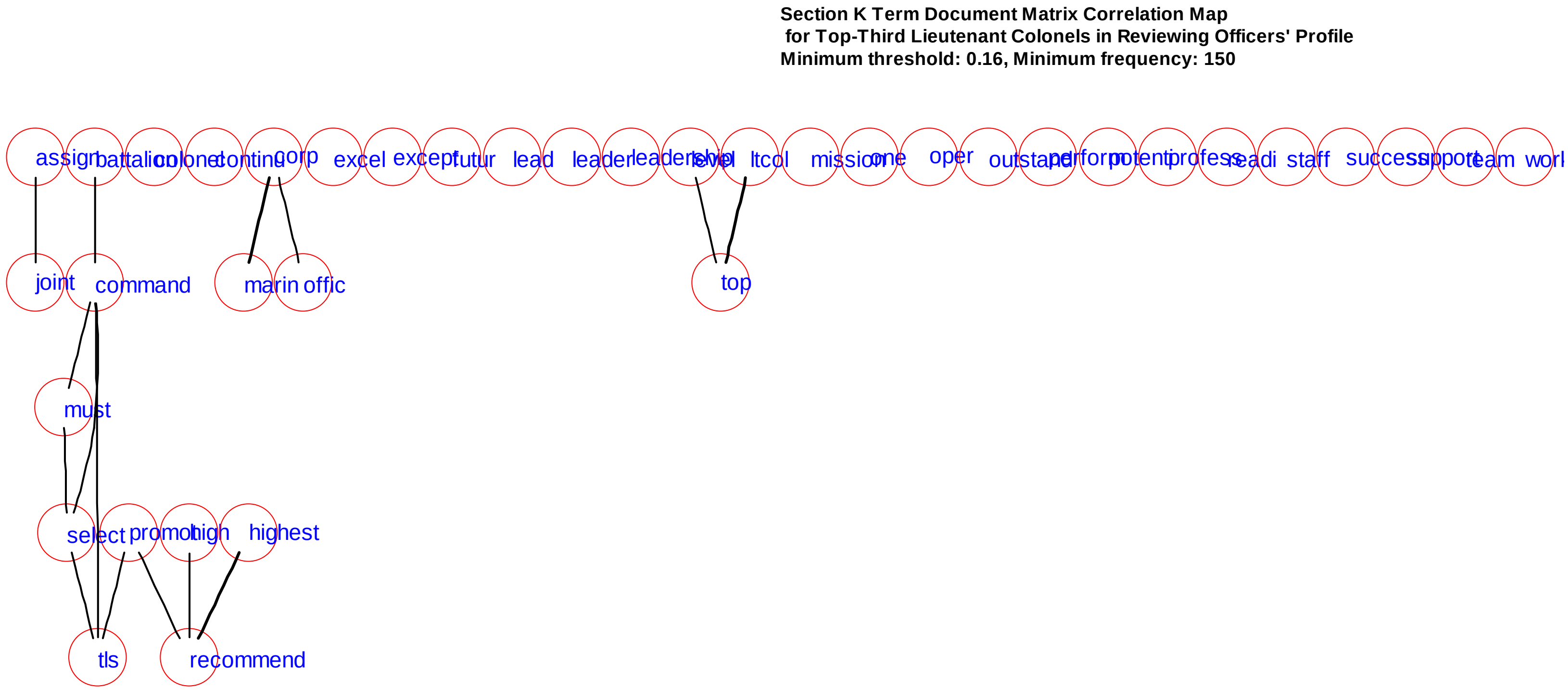


Figure 68. Unsupervised Correlation Map for Top Third LtCol Section K

APPENDIX D. SUPERVISED WORD CORRELATION PLOTS

A. REPORTING SENIOR

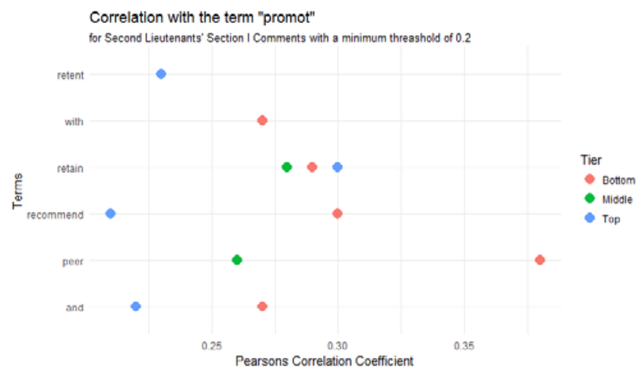


Figure 69. Correlation with the Word "promote" for 2ndLt Section I

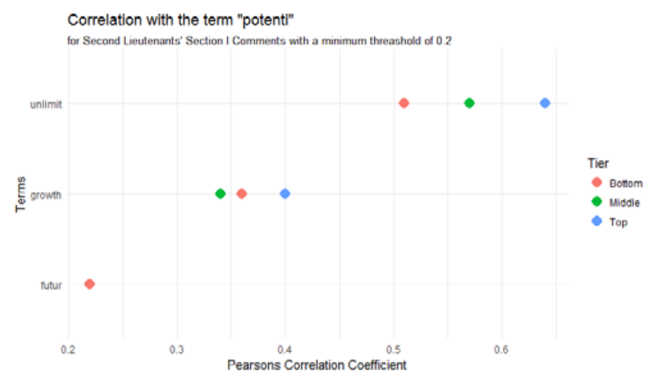


Figure 70. Correlation with the Word "potential" for 2ndLt Section I

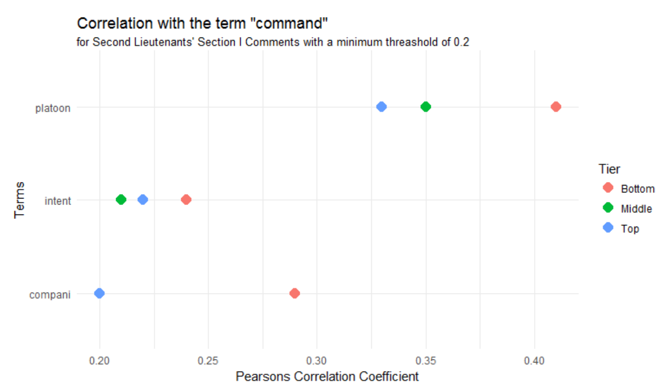


Figure 71. Correlation with the Word "command" for 2ndLt Section I

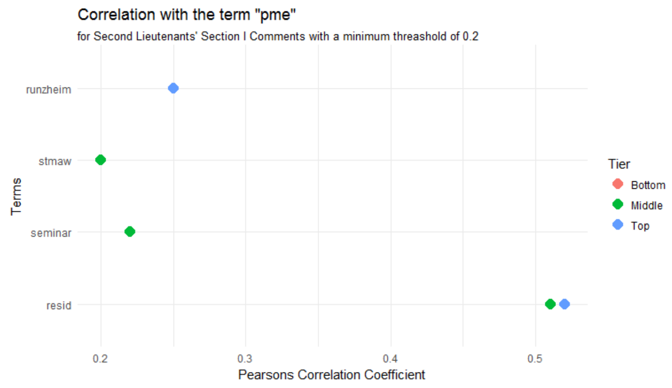


Figure 72. Correlation with the Word "pme" for 2ndLt Section I

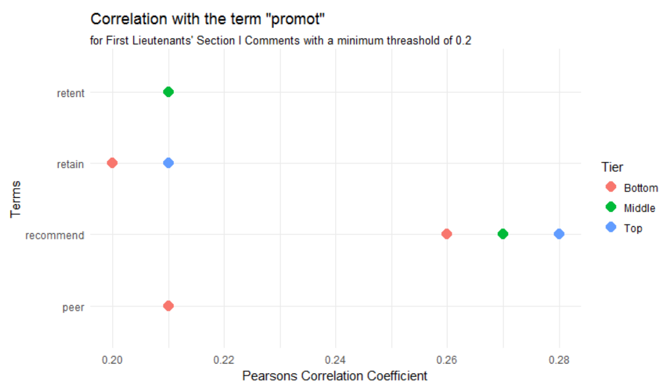


Figure 73. Correlation with the Word "promote" for 1stLt Section I

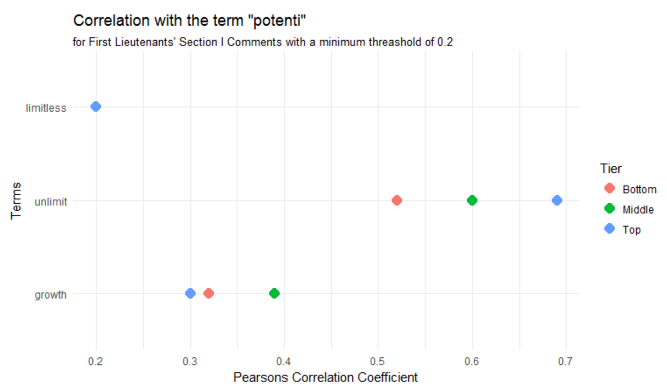


Figure 74. Correlation with the Word "promote" for 1stLt Section I

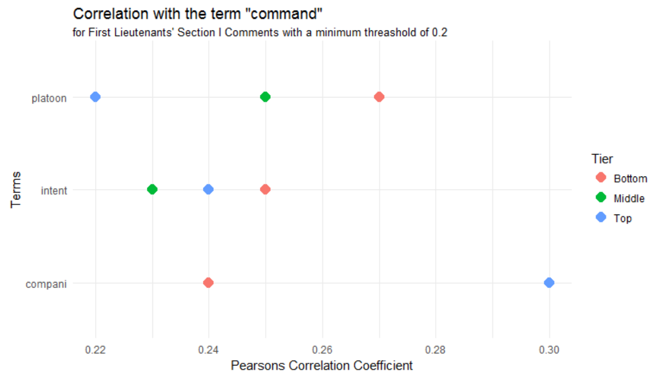


Figure 75. Correlation with the Word “command” for 1stLt Section I

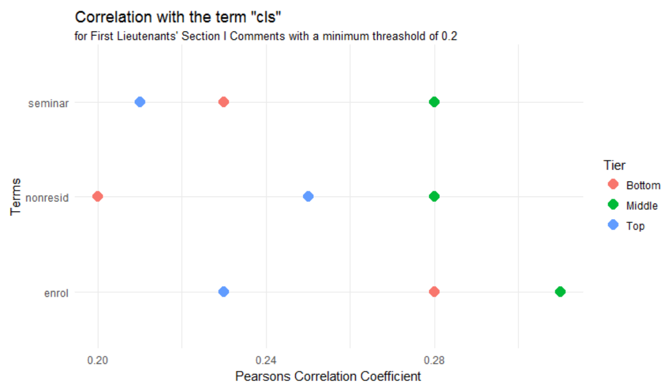


Figure 76. Correlation with the Word “cls” for 1stLt Section I

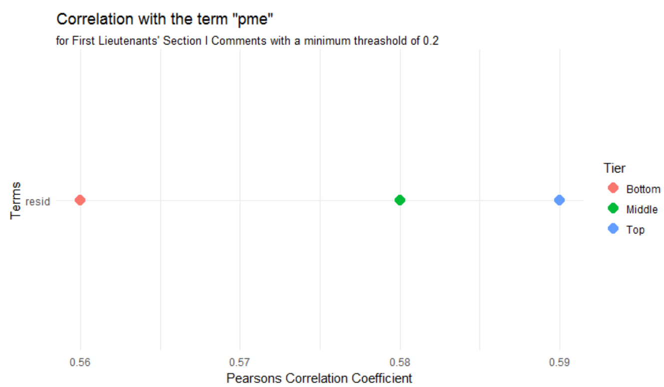


Figure 77. Correlation with the Word “pme” for 1stLt Section I

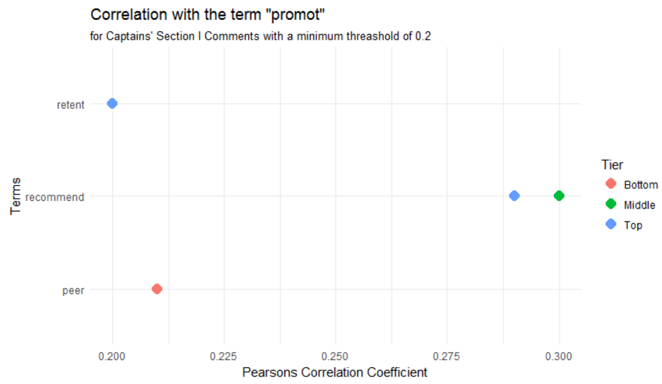


Figure 78. Correlation with the Word “promote” for Capt Section I

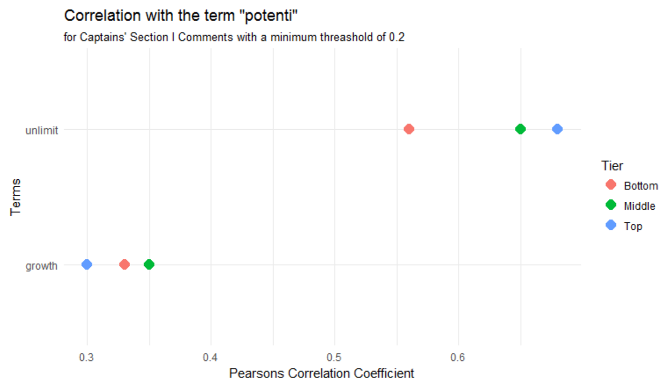


Figure 79. Correlation with the Word “potential” for Capt Section I

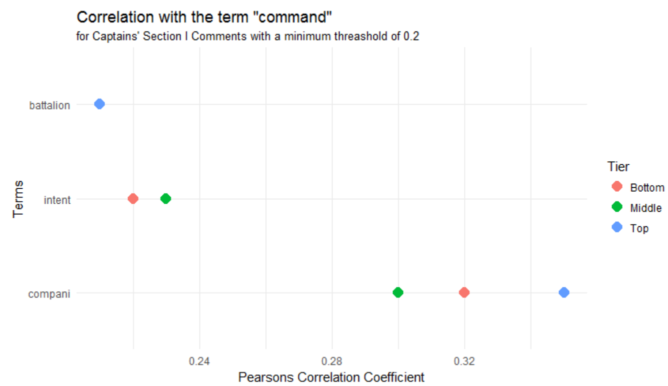


Figure 80. Correlation with the Word “command” for Capt Section I

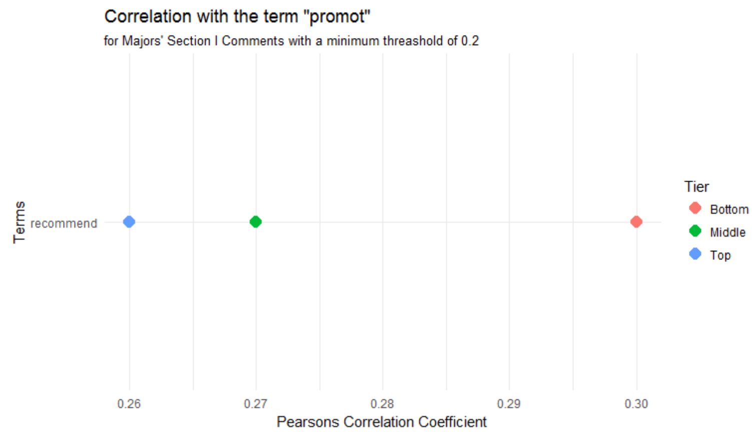


Figure 81. Correlation with the Word “promote” for Maj Section I

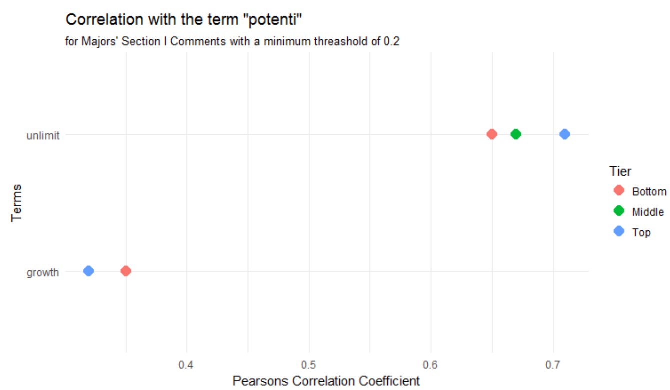


Figure 82. Correlation with the Word “potential” for Maj Section I

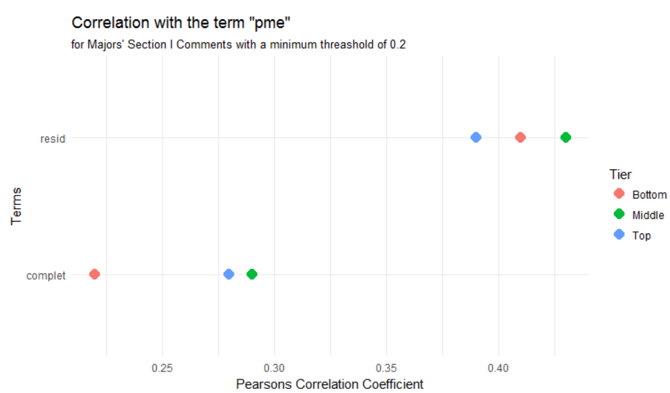


Figure 83. Correlation with the Word “pme” for Maj Section I

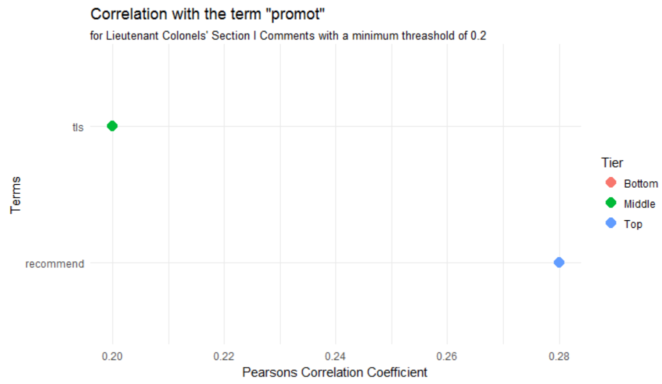


Figure 84. Correlation with the Word "promote" for LtCol Section I

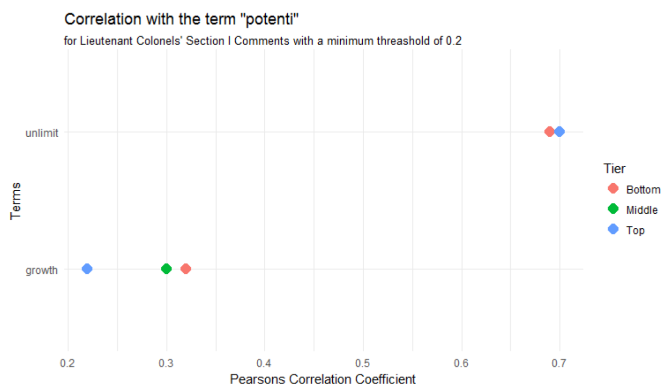


Figure 85. Correlation with the Word "potential" for LtCol Section I

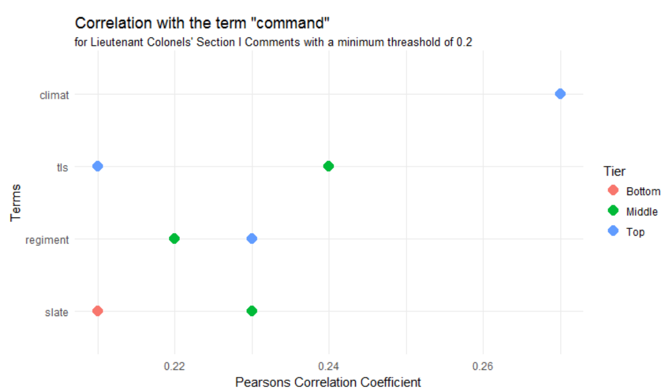


Figure 86. Correlation with the Word "command" for LtCol Section I

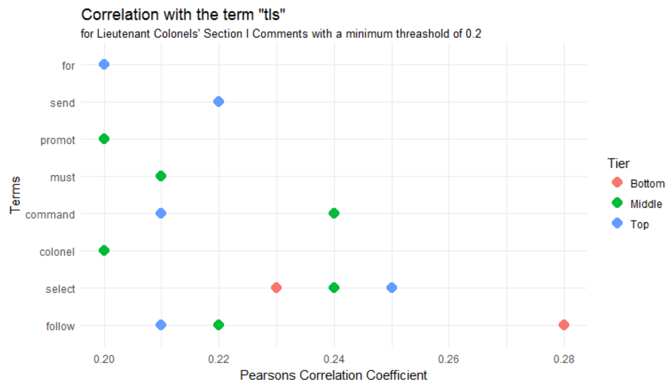


Figure 87. Correlation with the Word "tls" for LtCol Section I

B. REVIEWING OFFICER

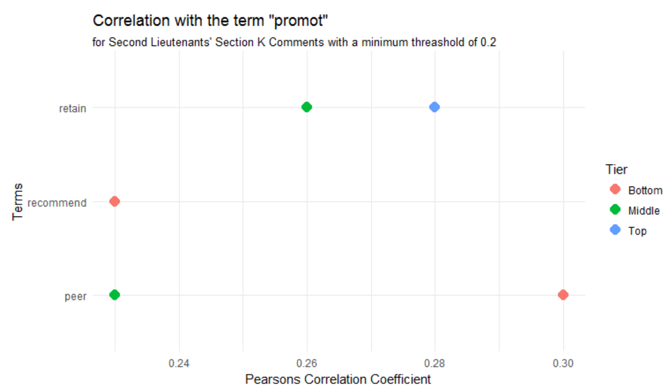


Figure 88. Correlation with the Word "promote" for 2ndLt Section K

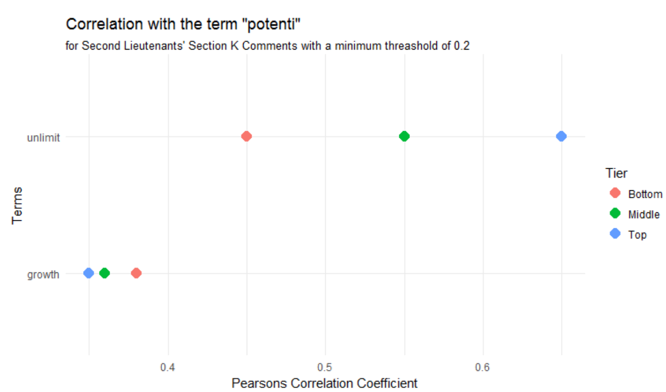


Figure 89. Correlation with the Word "potential" for 2ndLt Section K

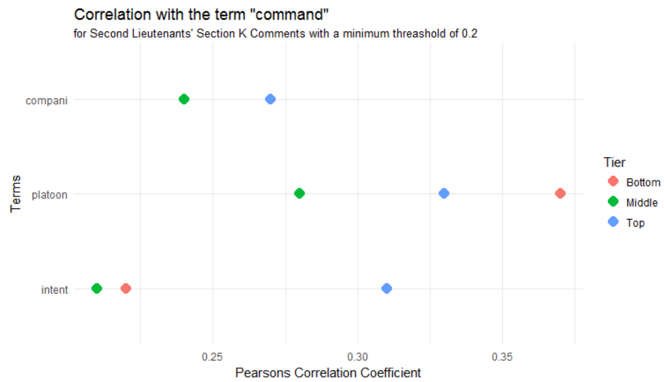


Figure 90. Correlation with the Word "command" for 2ndLt Section K

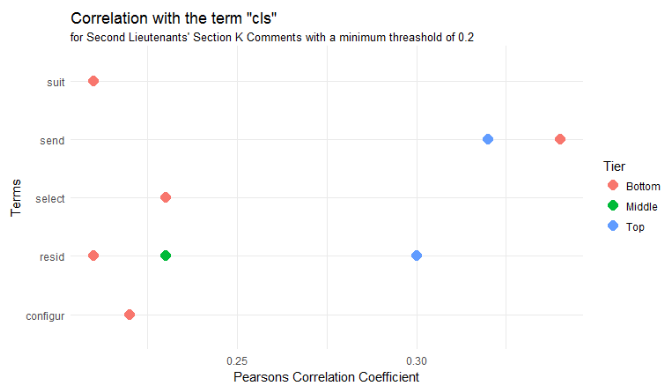


Figure 91. Correlation with the Word "cls" for 2ndLt Section K

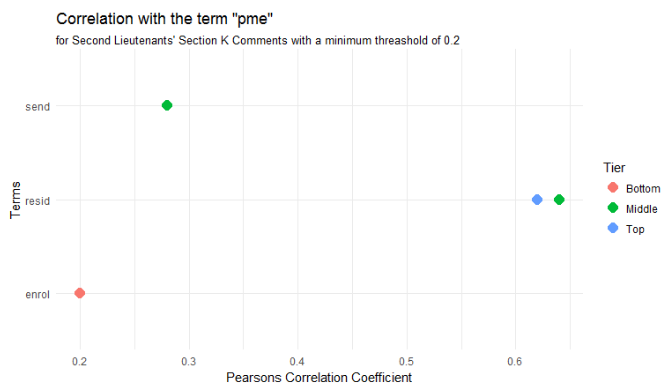


Figure 92. Correlation with the Word "pme" for 2ndLt Section K

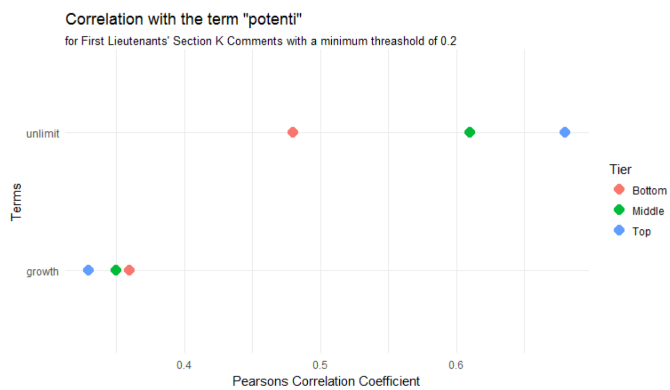


Figure 93. Correlation with the Word “potential” for 1stLt Section K

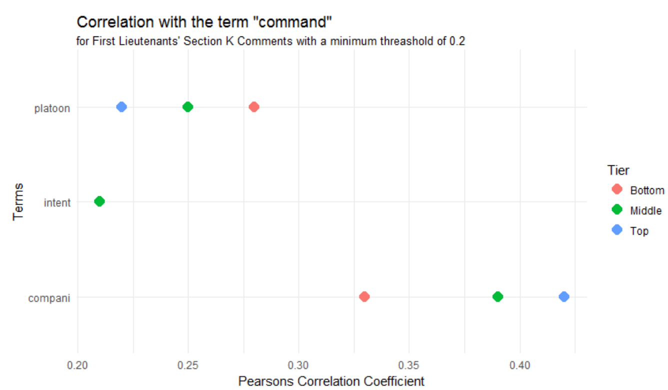


Figure 94. Correlation with the Word “command” for 1stLt Section K

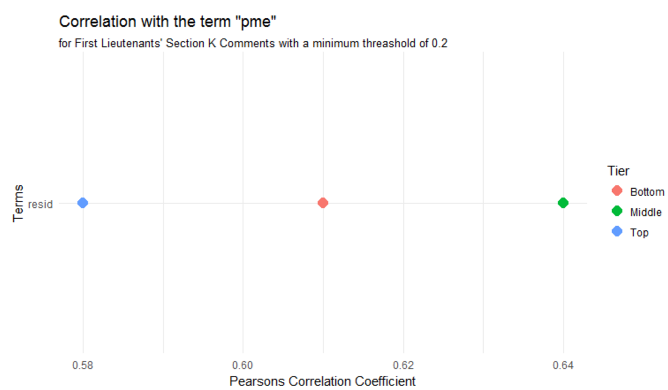


Figure 95. Correlation with the Word “pme” for 1stLt Section K

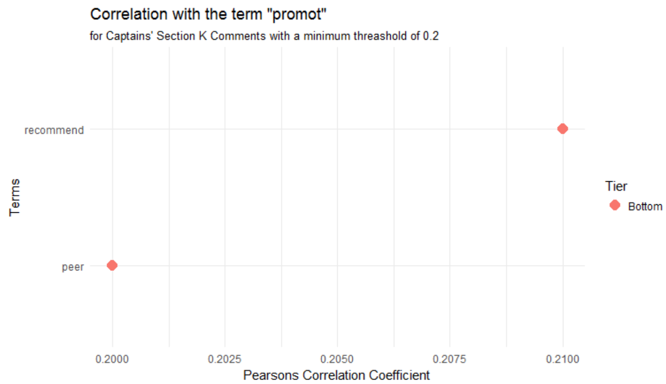


Figure 96. Correlation with the Word “promote” for Capt Section K

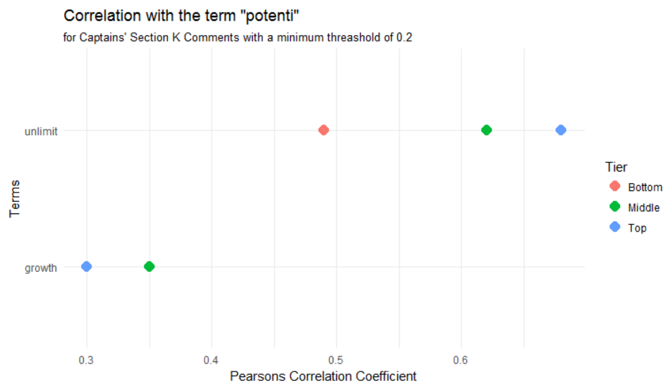


Figure 97. Correlation with the Word “potential” for Capt Section K

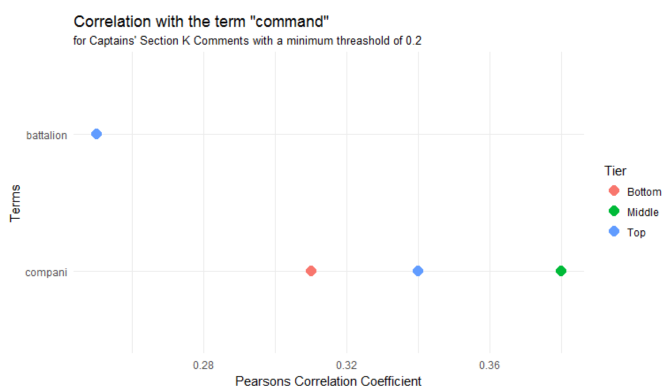


Figure 98. Correlation with the Word “command” for Capt Section K

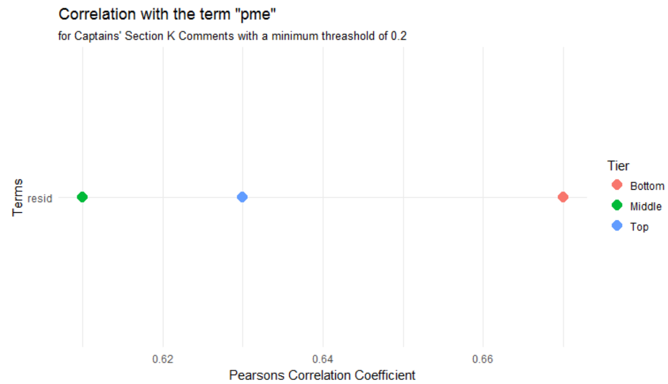


Figure 99. Correlation with the Word “pme” for Capt Section K

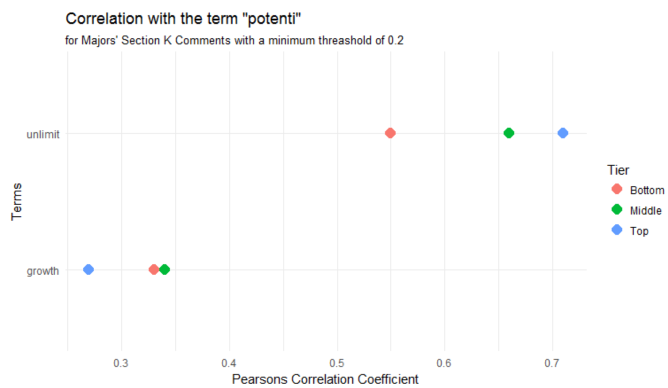


Figure 100. Correlation with the Word “potential” for Maj Section K

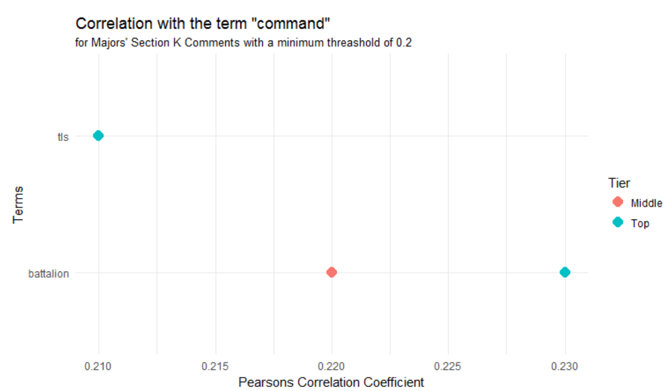


Figure 101. Correlation with the Word “command” for Maj Section K

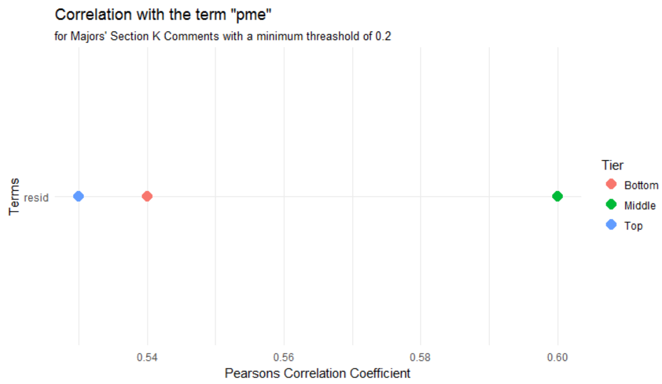


Figure 102. Correlation with the Word “pme” for Maj Section K

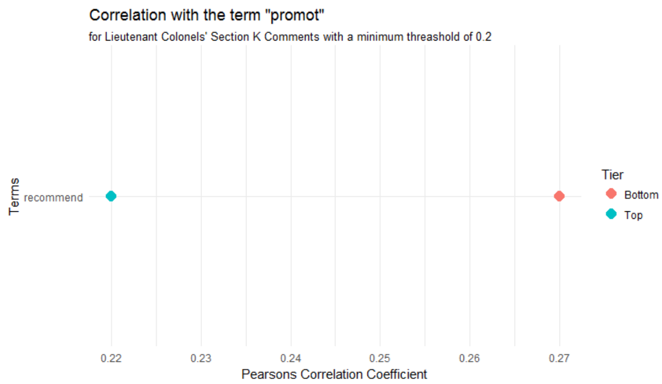


Figure 103. Correlation with the Word “promote” for LtCol Section K

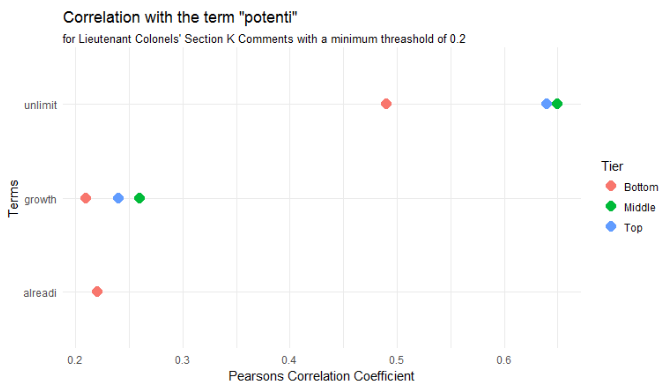


Figure 104. Correlation with the Word “potential” for LtCol Section K

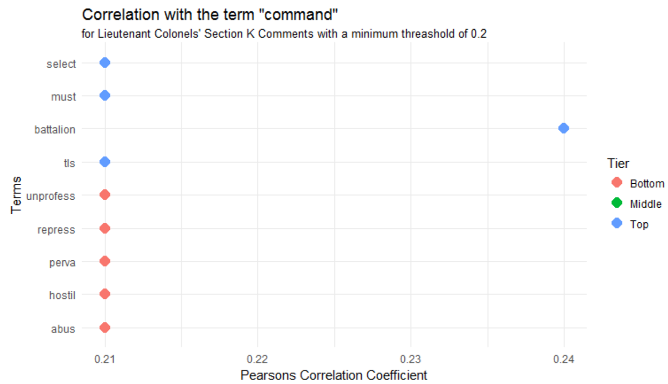


Figure 105. Correlation with the Word “command” for LtCol Section K

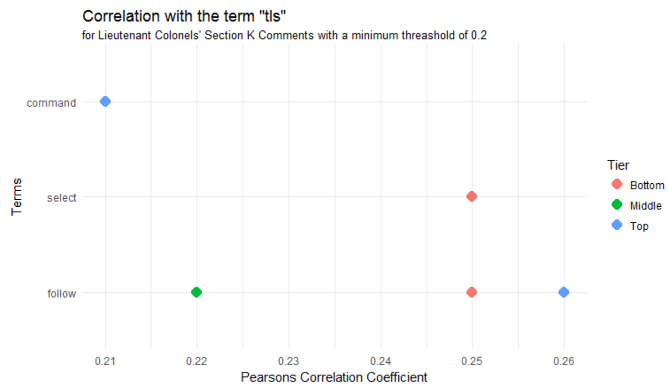


Figure 106. Correlation with the Word “tls” for LtCol Section K

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX E. PREDICTIVE MODEL PERFORMANCE

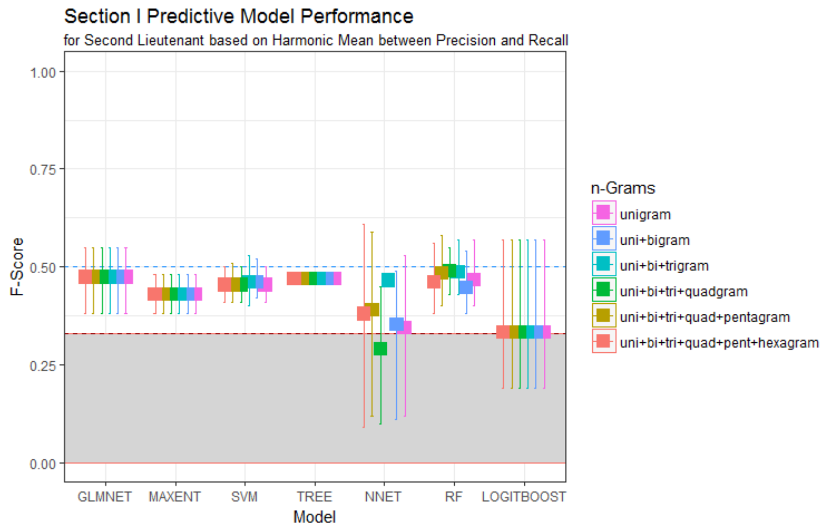


Figure 107. Section I Predictive Model Performance For Second Lieutenants

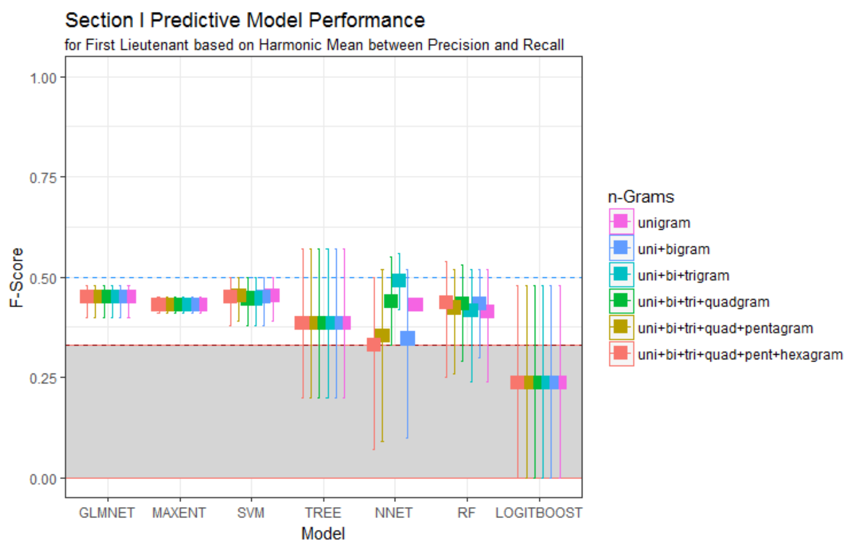


Figure 108. Section I Predictive Model Performance For First Lieutenants

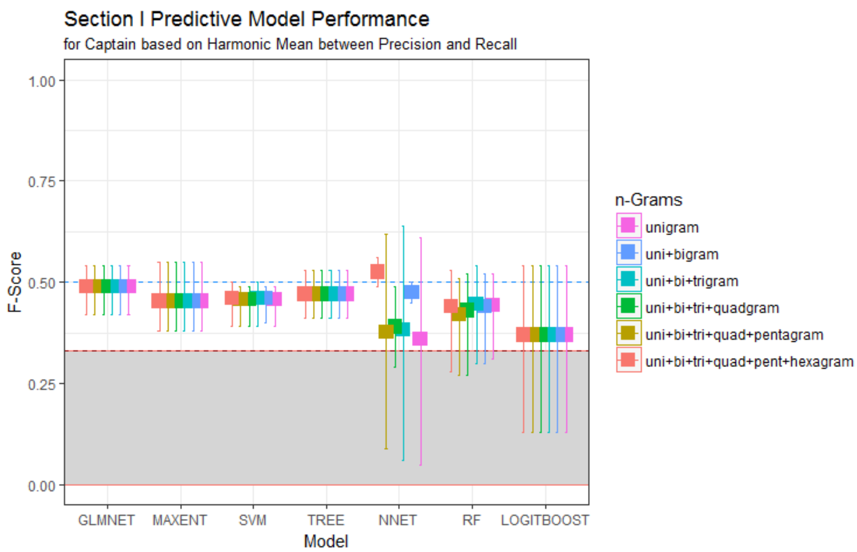


Figure 109. Section I Predictive Model Performance For Captains

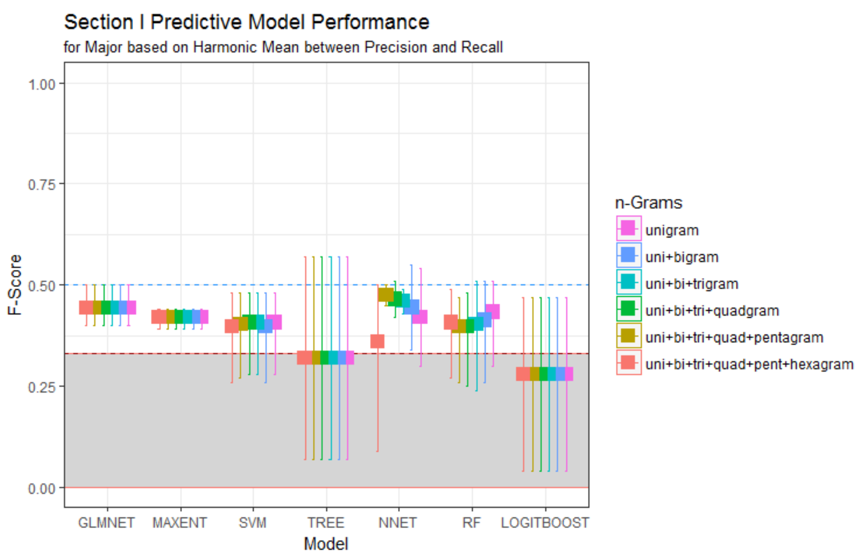


Figure 110. Section I Predictive Model Performance For Majors

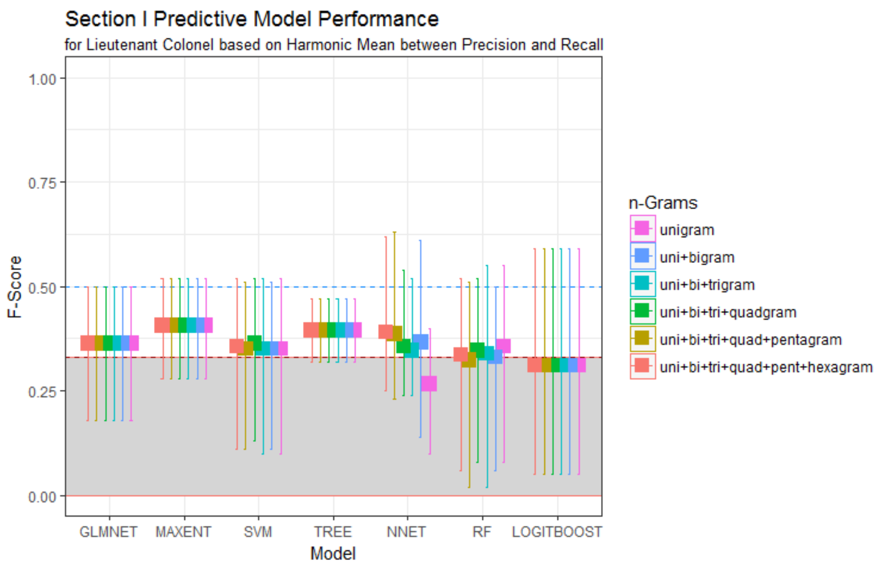


Figure 111. Section I Predictive Model Performance For Lieutenant Colonels

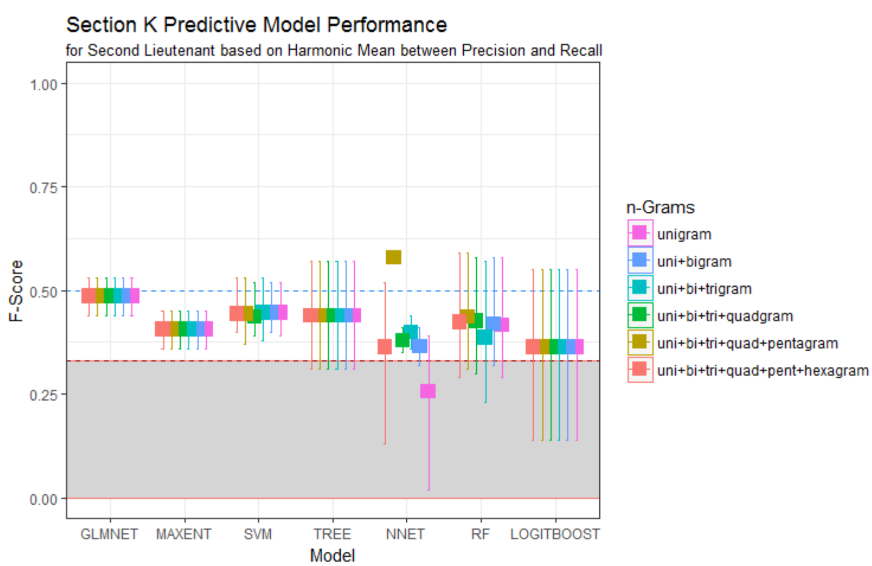


Figure 112. Section K Predictive Model Performance For Second Lieutenants

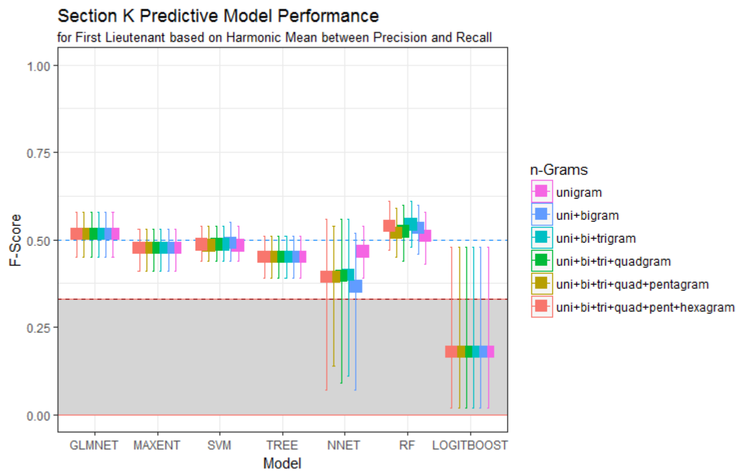


Figure 113. Section K Predictive Model Performance For First Lieutenants

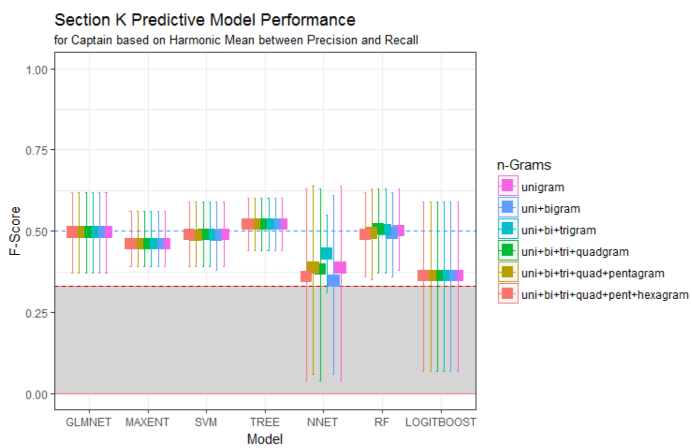


Figure 114. Section K Predictive Model Performance For Captains

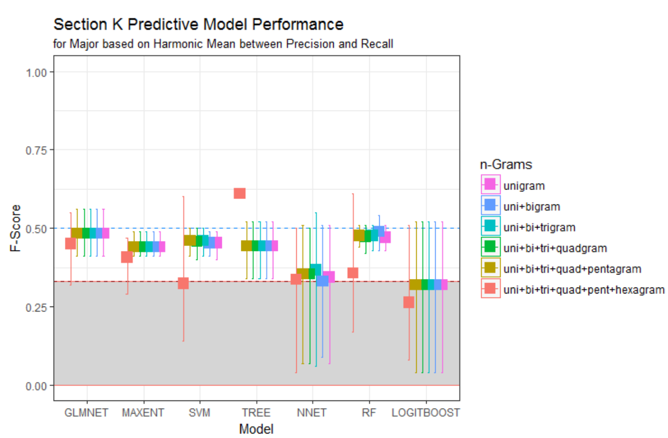


Figure 115. Section K Predictive Model Performance For Majors

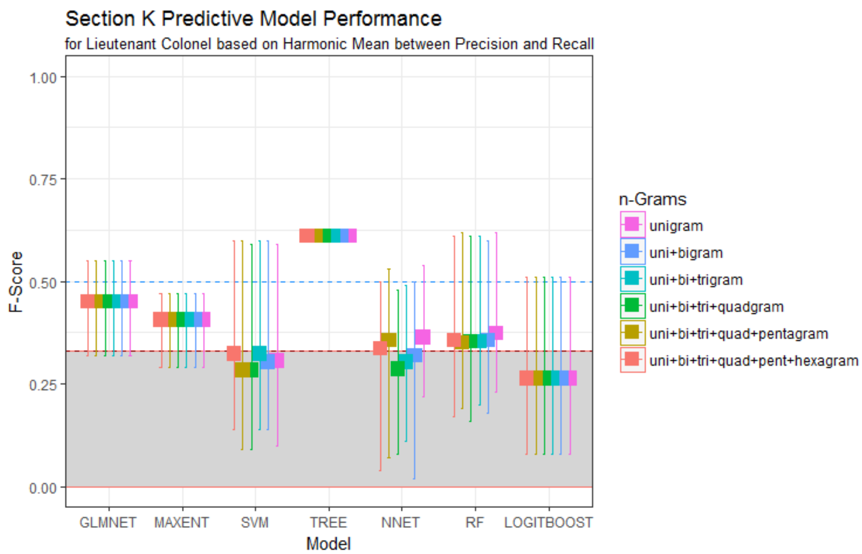


Figure 116. Section K Predictive Model Performance For Lieutenant Colonels

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Bach, S., & Sisson, K. (2000). *Performance Management: A Comprehensive Guide to Theory and Practice* (3rd ed.). Oxford, UK: Blackwell Publishers.

Benoit, K., & Nulty, P. (2016). *quanteda:Quantitative analysis of textual data.* Retrieved from R package version 0.9.8.5: https://CRAN.R-project.org/package=quanteda

Berger, A. (1996, July 5). *Maxent Modeling.* Retrieved from Carnegie Mellon University: http://www.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/node3.html

Bouchet-Val, M. (2014). *SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library.* Retrieved from R package version 0.5.1: https://CRAN.R-project.org/package=Snowball

Breiman, L. (2001). *Random Forest.* University of California, Berkeley, Statistic Department. Berkeley, CA: University of California, Berkeley.

Brownlee, J. (2016, February 8). *Machine Learning Mastery*. Retrieved from How to Build an Esemble of Machine Learning Algorithms in R: machinelearningmastery.com/machine-learning-ensembles-with-r/

Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mnining and Knowledge Discovery*, 121–167.

Buttrey, S. (2016a). *Ensemble Modeling.* Monterey, CA: Naval Postgraduate School.

Buttrey, S. (2016b). *Ensembles.* Monterey, CA: Naval Postgraduate School.

Chmiel, N. (2000). *Introduction to Work and Organizational Behavior.* Oxford, UK: Blackwell Publishers.

Chuang, J., Manning, C., & Heer, J. (2012, October). "Without the Clutter of Unimportant Words": Descriptive Keyphrases for Text Visualization. *ACM Transactions on Computer-Human Interaction, 19*(3), 1–29.

Clemens, A., Malone, L., Phillips, S., & Lee, G. (2012). *An Evaluation of the Fitness Report System for Marine Officers.* Arlington: Center for Naval Analyses.

Commandant of the Marine Corps . (2006). *Marine Corps Order P1400.31B Subject: Marine Corps Promotion Manual, Volume 1, Officer Promotion.* Washington, D.C.: Headquarters Marine Corps.

Commandant of the Marine Corps. (2013). *Marine Corps Order 1200.17E Military Occupational Specialties Manual.* Quantico, VA: United States Marine Corps. Retrieved January 12, 2016, from http://www.marines.mil/Portals/59/MCO%201200.17E.pdf

Commandant of the Marine Corps. (2015). *Marine Corps Order P1610.7: Performance Evaluation System.* Washington, D.C.: Headquarters Marine Corps.

Cornel University. (2004). *SMART Information Retrieval System.* Retrieved November 22, 2016, from http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stoplist/english.stop

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. *WWW '03 Proceedings of the 12th International Conference on World Wide Web* (pp. 519–528). Budapest, Hungary: Association for Computing Machinery.

Department of Defense. (2015). *United States Marine Corps: America's Expeditionary Force in Readiness.* Retrieved September 29, 2016, from http://www.marines.mil/Portals/59/Docs/MarineCorps_101.pdf

Dodd, J. (2016). *Fitness Reports* . The Basic School, Basic Officer Course. Quantico,VA: United States Marine Corps.

Drewry, D. (1998). *Fitness Report Writing Guide for Marines. Third Edition.* Pensacola, FL: Professional Management Spectrum, Inc.

DuBay, W. H. (2004). *The Principles of Readability.* Costa Mesa, California: Impact Information.

Ergun, L. (2003). *An Analysis of Officer Accession Programs and the Career Development of U.S. Marine Corps Officers.* Monterey, CA: Naval Postgraduate School.

Evans, D., Klavans, J., & Wacholder, N. (2000). Document processing with LinkIT. *Recherche d'Informations Assistee par Ordinateur*.

Evans, J. (1996). *Straightforward statistics for the behavioral sciences.* Pacific Grove, CA: Books/Cole Publishing.

Feinere, I., & Hornik, K. (2015). *tm: Text Mining Package.* Retrieved from R package version 0.6-2: https://CRAN.R-project.org/package=tm

Feinerr, I., & Hornik, K. (2015, July 3). *Text mining package "tm," Version 0.6-2*. Retrieved November 23, 2016, from https://cran.r-project.org/web/packages/tm/tm.pdf

Fellows, I. (2014). *wordcloud: Word Clouds.* Retrieved from R package version 2.5: https://CRAN.R-project.org/package=wordcloud

Freund, Y., & Schapire, R. (1999). A Short Introduction to Boosting. *Journal of Japanese Society for Artifical Intelligence, 14*(5), 771–780.

Freund, Yoay & Shapire, Robert. (1995). A Decision-Theoretic Generalization of on-line Learning and an Application of Boosting. AT&T Bell Laboratories. Murry Hill, NJ.

Friedlein, J. (2016). *The Islamic State Battle Plan: Press Release Natural Language Processing.* Monterey, CA: Naval Postgraduate School.

Friedman, J., Hastie, T., Simon, N., & Tibshirani, R. (2016). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.* Retrieved from CRAN.r-project.org: https://CRAN.R-project.org/package=glmnet

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1498–1512.

Gonzales, M. D. (2012). Analysis of promotion rates to lieutenant colonel and selection for command for USMC Aviation Supply and Maintenance Officers. *Masters Thesis*. Monterey, CA: Naval Postgraduate School.

Hansen, K., Gentry, J., Long, L., Gentleman, R., Falcon, S., Hahne, F., & Sarkar, D. (2016). *Rgraphviz: Provides plotting capabilities for R graph objects.* Retrieved from R package version 2.10: https://CRAN.R.project.org/package=Rgraphviz

Hastie, T., & Qian, J. (2014, June 26). *Glmnet Vignette*. Retrieved from Stanford: https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Machine Learning: Data Mining, Inference, and Prediction.* Standford, CA: Springer.

Hennig, C. (2015). *fpc: Flexible Procedures for Clustering.* Retrieved from R package version 2.1-10: https://CRAN.R-project.org/package=fpc

Henningsen, A. (2014). *micEcon: Microeconomic Analysis and Modeling.* Retrieved from R package version 0.6-12: https://CRAN.R-project.org/package=micEcon

Hornick, K., Buchta, C., & Zeileis, A. (2009). Open-Source Machine Learning: R Meets Weka. *Computational Statistics*, 225–232. Retrieved from http://doi.org/10.1007/s00180-008-0119-7

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning:R meets weka. *Computational Statistics*, *24*(2), 225–232.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 216–223.

Jobst, M., & Palmer, J. (2005). *Analysis of the USMC FITREP: Contemporary or Inflexible?* Monterey, CA: Naval Postgraduate School.

Jordan, D. W. (2011). *Re-Thinking Student Written Comments in Course Evaluations: Text Mining Unstructured Data for Program and Institutional Assessment.* Stanislaus, CA: California State University.

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & Atteveldt, W. v. (2012). *RTextTools: Automatic Text Classification via Supervised Learning.* Retrieved from CRAN: http://CRAN.R-project.org/package=RTextTools

Kincaid, P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of New Readability Formulas for Navy Enlisted Personnel.* Millington, TN: Chief of Naval Technical Training.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2016). *cluster:Cluster Analysis Basics and Extensions.* Retrieved from R package version 2.0.5: https://CRAN.R-project.org/package=cluster

Mahoney, E. (2016, October 26). MMRP 1: Officer Promotions. (P. Rigaut, Interviewer)

Marucci, D. (2016, December 9). History of the PES. (P. Rigaut, Interviewer)

McCullough, W. S., & Pitts, W. H. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.

Mckechnie, T. (2017). *Benefits for categorical data sets using tree distances and parallel processing.* Monterey, CA: Naval Postgraduate School.

McLaughlin, G. (1969). SMOG grading - a new readability formula. *Journal of Reading 22*, 639–646.

Michalke, E. (2016). *Using the koRpus Package for Text Analysis.* CRAN-R. Retrieved from https://cran.r-project.org/web/packages/koRpus/vignettes/koRpus_vignette.pdf

Michallke, M. (2016). *koRpus: An R Package for Text Analysis* . Retrieved from R package 0.06-5: https://CRAN.R-project.org/package=koRpus

Moore, M. (2017, January 11). Basic Officer Course Fitness Report Education. (P. Rigaut, Interviewer)

Naval Postgraduate School. (2017). *High Perfomance Computing*. Retrieved from Naval Postgraduate School: https://my.nps.edu/web/technology/hpc

Neller, R. (2016). *Fragmentary Order 01/2016: Advance to Contact.* Quantico, VA: Office of the Commandant.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of EMNLP*, 79–86.

Phillips, S., & Clemens, A. (2011). *The Fitness Report System for Marine Officers: Prior Research.* Arlington, VA: Center for Naval Analyses.

Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to ROC. *Journal of Machine Learning Techniques, 2*(1), 37–63.

R Core Team. (2016). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/

Reck, R. P., & Reck, R. A. (2007). *Generating and Rendering Readability Scores for Project Gutenberg Texts.* Project Gutenberg. Retrieved from http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/120Paper.pdf

Reynolds, J. L. (2011). Effect of Being an Aviator on Promotion Probability to O-5 in the USMC. *Masters Thesis*. Monterey, CA: Naval Postgraduate School.

Rinker, T. (2013). *qdap: Quantitative Discourse Analysis Package.* Retrieved from R package 2.2.5: http:/github.com/trinker/qdap

Ripleu, B., & Lapsley, M. (2016). *RODBC: ODBC Database Access.* Retrieved from R package version 1.3-14: https://CRAN.R-project.org/package=RODBC

Selivanov, D. (2016). *text2vec: Modern Text Mining Framework for R. R package version 0.4.0.* Retrieved from CRAN: https://CRAN.R-project.org/ package=text2vec

Seshan, V. E. (2016). *clinfun: Clinical Trial Design and Data Analysis Functions.* . Retrieved from R package version 1.0.13.: https://CRAN.R-project.org/ package=clinfun

Sprent, P., & Smeeton, N. (2007). *Applied Nonparametric Statistical Methods* (Fourth ed.). Boca Raton, FL: Chapman & Hall/CRC.

The Basic School. (2016a). *Discussion Group Academic Instructor Guide.* The Basic School, Basic Officer Course. Quantico, VA: United States Marine Corps.

The Basic School. (2016b). *Fitness Report Workshop.* The Basic School, Basic Officer Course. Quantico, VA: United States Marine Corps.

The Basic School. (2016c). Fitness Reports: B3K0477XQ-DM Student Handout. Quantico, VA: United States Marine Corps.

Therneau, T., Atkinson, B., & Ripley, B. (2015). *rpart: Recursive Partitioning and Regression Trees.* Retrieved from R package version 4.1-10: https://CRAN.R-project.org/package=rpart

United States Marine Corps. (1995). *How to Write a Fitness Report* (Vol. NAVMC 2794). Quantico, VA: United States Marine Corps.

Urbanek, S. (2016). rJava: Low-Level R to Java Interface. *R package version 0.9-8*. Retrieved from https://CRAN.R-project.org/package=rJava

Vryniotis, V. (2013, November 20). *DatumBox*. Retrieved from Machine Learning Tutorial: THe Max Entropy Text Classifier: http://blog.datumbox.com/machine-learning-tutorial-the-max-entropy-text-classifier/

Weiss, S., Indurkhya, N., & Zhang, T. (2010). *Fundatmentals of Predictive Text Mining.* New York, NY: Springer New York Inc. Retrieved November 22, 2016, from https://cran.r-project.org/web/packages/tm/tm.pdf

Wellner, B. (2017). *Naïve Bayes, Maximum Entropy and Text Classification.* Retrieved from Michtom School of Computer Science: http://www.cs.brandeis.edu/~cs134/TextClassification-NB-MaxEnt.pdf

Whitaker, L. (2017a). *Classification and Regression Trees.* Monterey, CA: Naval Postgraduate School.

Whitaker, L. (2017b). *Combining Models.* Monterey, CA: Naval Postgraduate School.

Whitaker, L. (2017c). *Fundamentals of Statistical Machine Learning.* Monterey, CA: Naval Postgraduate School.

Whitaker, L. (2017d). *Introduction to Statistical Machine Learning.* Monterey, CA: Naval Postgraduate School.

Whitaker, L. (2017). *Neural Network.* Monterey, CA: Naval Postgraduate School.

White, S. (2003). *The 2003 National Assessment of Adult Literacy.* Center for Education Statistics (NCES): U.S. Dept. of Education, Inst. of Education. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=

Wickam, h., Hester, J., & Francois, R. (2016). *readr: Read Tabular Data.* Retrieved from R package version 1.0.0: https://CRAN.R-project.org/package=readr

Wicklam, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Retrieved from https://CRAN.R-project.org/package=ggplot2

Wild, F. (2015). *Isa: Latent Semantic Analysis.* Retrieved from R package version 0.73.1: https://CRAN.R-project.org/package=lsa

Williams, G. (2011). *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*. Use R!: Springer.

Zhai, C. (2017). *Word Association Mining and Analysis.* Urbana, IL: University of Illinois at Urbana-Champaign.

Zive, B. (2015, March 25). Text Mining and N-Grams. *RPubs*, https://rpubs.com/ brianzive/textmining.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California